(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 November 2001 (08.11.2001)

PCT

(10) International Publication Number
WO 01/83782 A2

(51) International Patent Classification[7]: C12N 15/57, 9/64, C07K 16/40, C12Q 1/37, 1/68 // A61P 3/00, 9/00, 25/00, 29/00, 35/00, 37/00

(21) International Application Number: PCT/US01/14431

(22) International Filing Date: 4 May 2001 (04.05.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/201,879 4 May 2000 (04.05.2000) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/201,879 (CIP)
Filed on 4 May 2000 (04.05.2000)

(71) Applicant *(for all designated States except US)*: **SUGEN, INC.** [US/US]; 230 E. Grand Avenue, South San Francisco, CA 94080-4811 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **PLOWMAN, Gregory, D.** [US/US]; 35 Winding Way, San Carlos, CA 94070 (US). **WHYTE, David** [US/US]; 2623 Barclay Way, Belmont, CA 94002 (US). **SUDARSANAM, Sucha** [US/US]; 20 Corte Patencio, Greenbrae, CA 94904 (US). **MANNING, Gerard** [US/US]; 844 Fremont Street #4, Menlo Park, CA 94025 (US). **CAENEPEEL, Sean** [US/US]; 447 Orange Street #22, Oakland, CA 94610 (US). **PAYNE, Vilia** [—/US]; Saint Louis, MO (US).

(74) Agents: **BURROUS, Beth, A.** et al.; Foley & Lardner, 3000 K. Street, NW, Washington, DC 20007-5109 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU,

[Continued on next page]

(54) Title: NOVEL PROTEASES

(57) Abstract: The present invention relates to protease polypeptides, nucleotide sequences encoding the protease polypeptides, as well as various products and methods useful for the diagnosis and treatment of various protease-related diseases and conditions. Through the use of a bioinformatics strategy, mammalian members of the of PTK's and STK's have been identified and their protein structure predicted.

>SGPR_140_SEQ ID NO:1
ATGAGGGGCCTTGTGGTATTCCTTGCAGTCTTTGCTCTCTCTGAGGTCAATGCCATCACCAGGGTTC
CTCTGCACAAAGGGAAGTCGCTGAGGAGGGCCCTGAAGGAGCGCAGGCTCCTGGAGGACTTCCTG
AGGAATCACCATTATGCAGTCAGCAGGAAGCACTCCAGCTCTGGGGTGGTGGCCAGCGAGTCTCT
GACCAACTACCTGGATTGTCAGTACTTTGGGAAGATCTACATCGGGACCCTTCCCCAGAAGTTCAC
CTTGGTGTTTGATACAGGCTCCCCGGATATCTGGGTGCCCTCTGTCTACTGCAACAGTGATGCCTGT
CAGAACCACCAACGCTTCGATCCGTCCAAGTCCTCCACCCAGAACATGGGCAAGTCCTCGTCCATC
CAGTATGGCACAGGCAGCATGCGGGCTTGCTGGGCTATGACACTGTCACCGTCTCCAACATTGTG
GACCCCACCAGACTGTGGGTCTGAGCACCCAGGAACCTGGCGACGTCTTCACCTACTCCGAGTTT
GATGGATCCTGGGGCTGGCCTATCCCTCTCTTGCCTCTGAGTACGCGCTGCGCCTTGGTTTCAGG
AATGACCAGGGGAGCATGCTCACGCTGAGGGCCATTGATCTGTCGTACTACACAGGCTCCCTGCAC
TGGATACCCATGACTGCAAGAATACTGGCAGTTCACTGTGGACAGGAAGGACCTGGGGAGGGAGG
GCTGGATGAGGCCATCTTGCATACCTTTGGAAGTGTCATCATTGACGGCGTGGTGGTGGCCTGTGA
CGGTGGCTGTCAGGCCATCCTGGACACCGGCACCTCCCTGCTGGTGGGGCCTGGTGGCAACATCCT
CAACATCCAGCAGGCCATTGGACGCACTGCGGGCCAGTACAATGAGTTTGACATCGACTGCGGGC
GCCTGAGCAGCATTCCCACGGCTGTCTTCGAGATCCACGGCAAGAAGTACCCCCTGCCACCCTCCG
CCTATACCAGCCAGGACCAGGGCTTCTGCACCAGTGGTTTCCAGGGTGACTATAGTTCCCAGCAGT
GGATCCTGGGGAATGTCTTCATCTGGGAGTATTACAGTGTCTTTGACAGGACCAATAACCGTGTGG
GGCTGGCGAAGGCTGTCTGA

>SGPR_197_SEQ ID NO:2
ATGGATAGATGCAAACATGTAGGGCGGTTACGGCTCGCCCAGGACCACTCCATCCTGAACCCTCA
GAAGTGGTGCTGCTTAGAGTGTGCCACCACCGAGTCCGTGTGGGCCTGCCTCAAGTGCTCCCACGT
GGCCTGCGGCCGCTATATTGAGGACCACGCCCTGAAACACTTTGAGGAGACGGGACACCCGCTAG
CCATGGAAGTCCGGGATCTCTACGTGTTCTGTTACCTGTGCAAGGACTACGTGCTCAATGATAACC
CAGAGGGGGACCTGAAGCTGCTAAGAAGCTCCCTCCTGGCGGTCCGGGGCCAGAAACAGGACACG
CCGGTGAGACGTGGGCGGACGCTGCGGTCCATGGCTCCGGTGAGGACGTGGTCCTGCCGCGCACG
CGCTCCTCAGGGACAGCCGCAGATGCTCCGCTCTGTGGTACCGGCGTCAGCGCCTGCTGGCCAG
GACGCTGCGCCTGTGTTCGAGAAGAGCTCCCGGGGCCAGGCGAAGCTGGAGCAGCGGCGGCAG
GAGGAGGCCCTGGAGCGCAAGAAGGAGGAGGCCGCGGAGGCGGCGGCGCGAGCCGGCCATGGCCC
CAGGCGTCACGGGCCTGCGCAACCTGGGCAACACCTGCTACATGAACTCCATCCTCCAGGTGCTCA
GCCACCTCCAGAAGTTCCGAGAATGTTTCCTCAACCTTGACCCTTCCAAAACGGAACATCTGTTTC
CCAAAGCCACCAACGGGAAGACTCAGCTTTCTGGCAAGCCAACCAACAGCTCGGCCACGGAGCTG
TCCTTGAGAAATGACAGGGCCGAGGCATGCGAGCGGGAGGGTTTCTGCTGGAACGGCAGGGCCTC
CATTAGTCGGAGTCTGGAGCTCATCCAGAACAAGGAGCCGAGTTCAAAGCACATTTCCCTCTGCCG
TGAACTGCACACCCTCTTCCGAGTCATGTGGTCCGGGAAGTGGGCCCTAGTGTCGCCCTTCGCCAT
GCTCCACTCAGTGTGGAGCCTGATCCCTGCCTTCCGCGGCTACGACCAACAGGACGCGCAGGAATT
TCTCTGCGAGCTGCTGCACAAGGTGCAGCAGGAACTCGAGTCTGAGGGCACCACACGCCGGATCC
TCATCCCCTTCTCCCAGAGGAAGCTCACCAAACAGGTCTTAAAGGTGGTGAATACCATATTTCATG
GGCAGCTGCTCAGTCAGGGAAGGTGGTCTGGCCGTAATCATCGAGAGAAGATTGGGGTCCATGTC
GTCTTTGACCAGGTATTAACCATGGAACCTTACTGCTGCAGGGACATGCTCTCCTCTCTTGACAAA
GAGACCTTTGCCTATGATCTCTCCGCAGTGGTCATGCATCACGGGAAAGGGTTTGGCTCAGGACAC
TACACAGCCTATTGCTACAACACAGAGGGAGGGGAGCAGACCCAGGGTTTGGCCATCACCAACCG
GGAGTACGGCCTAAGCCAGAGGGAGCTGGCACCACCTTCGAAAGCATTCCCTTTGATGTGA

>SGPR_005_SEQ ID NO:3
ATGGGGCCAAGACTCATTCCGTTTCTATTTTTGTTTGTTTACCCTATTCTCTGCAGGATCATTCTGA
GGAAAGGCAAGTCTATCCGGCAGAGAATGGAGGAGCAGGGTGTACTGGAGACGTTTCTGAGGGAC
CACCCAAAGGCTGATCCAATTGCCAAGTATTATTTCAATAATGATGCTGTTGCTTATGAGCCTTC
ACCAACTACCTGGATTCTTTCTACTTTGGGGAGATCAGCACTGGGACACCACCCCAAAATTTCCTA
GTCTCTTTGATACGGGTTCCTCCAATCTGTAGCCTGCCCTCCATCTACTGCCAGAGCCAAGTCTGCT
CCAATCACAACAGGTTCAATCCCAGCCTGTCCTCCACCTTCAGAAACGATGGACAAACCTATGGAC
TATCCTATGGGAGTGGCAGCCTGAGTGTGTTCCTGGGCTATGACACTGTGACTGTTCATAACATCG
TTGTCAATAACCAGGAGTTTGGCCTGAGTGAGAATGAGCCCAGCGACCCCTTTTACTATTCAGACT
TTGACGGGATCCTGGGAATGGCCTACCCAAACATGGCAGAGGGGAATTCCCCTACAGTAATGCAG
GGGATGCTGCAGCAGAGCCAGCTTACTCAGCCCGTCTTCAGCTTCTACTTCACCTGCCAGCCAACC
CGCCAGTATTGTGGAGAGCTCATCCTTGAGGTGTGGACCCCAACCTTTATTCTGGTCAGATCATC
TGGACCCCTGTCAGCCCGGAACTGTACTGGCAGATTGCCATCGAGGAATTTGCCATCGGTAACCAG

CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

NOVEL PROTEASES

The present invention claims priority to provisional application serial no. 60/201,879, filed May 4, 2000, which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

The present invention relates to protease polypeptides, nucleotide sequences encoding the protease polypeptides, as well as various products and methods useful for the diagnosis and treatment of various protease-related diseases and conditions.

BACKGROUND OF THE INVENTION

Proteases and Human Disease

"Protease," "proteinase," and "peptidase" are synonymous terms applying to all enzymes that hydrolyse peptide bonds, *i.e.* proteolytic enzymes. Proteases are an exceptionally important group of enzymes in medical research and biotechnology. They are necessary for the survival of all living creatures, and are encoded by 1-2% of all mammalian genes. Rawlings and Barrett (MEROPS: the peptidase database. *Nucleic Acids Res.*, 1999, 27:325-331) (http://www.babraham.co.uk/Merops/Merops.htm (Which is incorporated herein by reference in its entirety including any figures, tables, or drawings.)) have classified peptidases into 157 families based on structural similarity at the catalytic core sequence. These families are further classed into 26 clans, based on indications of common evolutionary relationship. Peptidases

play key roles in both the normal physiology and disease-related pathways in mammalian cells. Examples include the modulation of apoptosis (caspases), control of blood pressure (renin, angiotensin-converting enzymes), tissue remodeling and tumor invasion (collagenase), the development of Alzheimer's Disease (β-secretase), protein turnover and cell-cycle regulation (proteosome), and inflammation (TNF-α convertase). (Barrett *et al.*, Handbook of Proteolytic Enzymes, 1998, Academic Press, San Diego which is incorporated herein by reference in its entirety including any figures, tables, or drawings.)

Peptidases are classed as either exopeptidases or endopeptidases. The exopeptidases act only near the ends of polypeptide chains: aminopeptidases act at the free N-terminus and carboxypeptidases at the free C-terminus. The endopeptidases are divided, on the basis of their mechanism of action, into six sub-subclasses: aspartyl endopeptidases (3.4.23), cysteine endopeptidases (3.4.22), metalloendopeptidases (3.4.24), serine endopeptidases (3.4.21), threonine endopeptidases (3.4.25), and a final group that could not be assigned to any of the above classes (3.4.99). (Enzyme nomenclature and numbering are based on "Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) 1992, (http://www.chem.qmw.ac.uk/iubmb/enzyme/EC34/intro.html).)

In serine-, threonine- and cysteine-type peptidases, the catalytic nucleophile is the reactive group of an amino acid side chain, either a hydroxyl group (serine- and threonine-type peptidases) or a sulfhydryl group (cysteine-type peptidases). In aspartic-type and metallopeptidases, the nucleophile is commonly an activated water molecule. In aspartic-type peptidases, the water molecule is directly bound by the side chains of aspartate residues. In metallopeptidases, one or two metal ions hold the water molecule in place, and charged amino acid side chains are ligands for the metal ions. The metal may be zinc, cobalt or manganese. One metal ion is usually attached to three amino acid ligands. Families of peptidases are referred to by use of the numbering system of Rawlings & Barrett (Rawlings, N. D. & Barrett, A. J. MEROPS: the peptidase database. *Nucleic Acids Research* 27 (1999) 325-331,

which is incorporated herein by reference in its entirety including any figures, tables, or drawings).

Protease Families

1. Aspartyl proteases (Prosite number PS00141)

Aspartyl proteases, also known as acid proteases, are a widely distributed family of proteolytic enzymes in vertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Enzymes in this class include cathepsin E, renin, presenilin (PS1), and the APP secretases.

2. Cysteine proteases (Prosite PDOC00126)

Eukaryotic cysteine proteases are a family of proteolytic enzymes which contain an active site cysteine. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. Peptidases in this family with important roles in disease include the caspases, calpain, hedgehog, ubiquitin hydrolases, and papain.

3. Metalloproteases (Prosite PDOC00129)

The metalloproteases are a class which includes matrix metalloproteases (MMPs), collagenase, stromelysin, gelatinase, neprylisin, carboxypeptidase, dipeptidase, and membrane-associated metalloproteases, such as those of the ADAM family. They require a metal co-factor for activity; frequently the required metal ion is zinc but some metalloproteases utilize cobalt and manganese.

Proteins of the extracellular matrix interact directly with cell surface receptors thereby initiating signal transduction pathways and modulating those triggered by growth factors, some of which may require binding to the extracellular

matrix for optimal activity. Therefore the extracellular matrix has a profound effect on the cells encased by it and adjacent to it. Remodeling of the extracellular matrix requires protease of several families, including metalloproteases (MMPs).

4. Serine proteases (S1) (Prosite PS00134 trypsin-his; PS00135 trypsin-ser)

The catalytic activity of the serine proteases from the trypsin family is provided by a charge relay system involving an aspartic acid residue hydrogen-bonded to a histidine, which itself is hydrogen-bonded to a serine. The sequences in the vicinity of the active site serine and histidine residues are well conserved in this family of proteases. A partial list of proteases known to belong to this large and important family include: blood coagulation factors VII, IX, X, XI and XII; thrombin; plasminogen; complement components C1r, C1s, C2; complement factors B, D and I; complement-activating component of RA-reactive factor; elastases 1, 2, 3A, 3B (protease E); hepatocyte growth factor activator; glandular (tissue) kallikreins including EGF-binding protein types A, B, and C; NGF-γ hain, γ-renin, and prostate specific antigen (PSA); plasma kallikrein; mast cell proteases; myeloblastin (proteinase 3) (Wegener's autoantigen); plasminogen activators (urokinase-type, and tissue-type); and the trypsins I, II, III, and IV. These peptidases play key roles in coagulation, tumorigenesis, control of blood pressure, release of growth factors, and other roles.

5. Threonine peptidases (T1) – (Prosite PDOC00326/PDOC00668)

Threonine proteases are characterized by their use of a hydroxyl group of a threonine residue in the catalytic site of these enyzmes. Only a few of these enzymes have been characterized thus far, such as the 20S proteasome from the archaebacterium *Thermoplasma acidophilum* (Seemuller *et al.*, 1995, *Science*, 268:579-82, and chapter 167 of Barrett *et al.*, Handbook of Proteolytic Enzymes, 1998, Academic Press, San Diego).

SUMMARY OF THE INVENTION

This invention concerns the isolation and characterization of novel sequences of human proteases. These sequences are obtained via bioinformatics searching strategies on the predicted amino acid translations of new human genetic sequences. These sequences, now identified as proteases, are translated into polypeptides which are further characterized. Additionaly, the nucleic acid sequences of these proteases are used to obtain full-length cDNA clones of the proteases. The partial or complete sequences of these proteases are presented here, together with their classification, predicted or deduced protein structure.

Modulation of the activities of these proteases will prove useful therapeutically. Additionally, the presence or absence of these proteases or the DNA sequence encoding them will prove useful in diagnosis or prognosis of a variety of diseases. In this regard, Example 8 describes the chromosomal localization of proteases of the present invention, and describes diseases mapping to the chromosomal locations of the proteases of the invention.

A first aspect of the invention features an identified, isolated, enriched, or purified nucleic acid molecule encoding a protease polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70The term "identified" in reference to a nucleic acid is meant that a sequence was selected from a genomic, EST, or cDNA sequence database based on being predicted to encode a portion of a previously unknown or novel protease.

By "isolated" in reference to nucleic acid is meant a polymer of 10 (preferably 21, more preferably 39, most preferably 75) or more nucleotides conjugated to each other, including DNA and RNA that is isolated from a natural source or that is synthesized as the sense or complementary antisense strand. In certain embodiments of the invention, longer nucleic acids are preferred, for example those of 300, 600, 900, 1200, 1500, or more nucleotides and/or those having at least 50%, 60%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to a sequence selected from the group consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35.

It is understood that by nucleic acid it is meant, without limitation, DNA, RNA or cDNA, and where the nucleic acid is RNA, the thymine (T) will be uracil (U).

The isolated nucleic acid of the present invention is unique in the sense that it is not found in a pure or separated state in nature. Use of the term "isolated" indicates that a naturally occurring sequence has been removed from its normal cellular (*i.e.*, chromosomal) environment. Thus, the sequence may be in a cell-free solution or placed in a different cellular environment. The term does not imply that the sequence is the only nucleotide chain present, but that it is essentially free (preferably about 90% pure, more preferably at least about 95% pure) of non-nucleotide material naturally associated with it, and thus is distinguished from isolated chromosomes.

By the use of the term "enriched" in reference to nucleic acid is meant that the specific DNA or RNA sequence constitutes a significantly higher fraction (2- to

5-fold) of the total DNA or RNA present in the cells or solution of interest than in normal or diseased cells or in the cells from which the sequence was taken. This could be caused by a person by preferential reduction in the amount of other DNA or RNA present, or by a preferential increase in the amount of the specific DNA or RNA sequence, or by a combination of the two. However, it should be noted that enriched does not imply that there are no other DNA or RNA sequences present, just that the relative amount of the sequence of interest has been significantly increased. The term "significant" is used to indicate that the level of increase is useful to the person making such an increase, and generally means an increase relative to other nucleic acids of about at least 2-fold, more preferably at least 5-fold, more preferably at least 10-fold or even more. The term also does not imply that there is no DNA or RNA from other sources. The DNA from other sources may, for example, comprise DNA from a yeast or bacterial genome, or a cloning vector such as pUC19. This term distinguishes from naturally occurring events, such as viral infection, or tumor-type growths, in which the level of one mRNA may be naturally increased relative to other species of mRNA. That is, the term is meant to cover only those situations in which a person has intervened to elevate the proportion of the desired nucleic acid.

It is also advantageous for some purposes that a nucleotide sequence be in purified form. The term "purified" in reference to nucleic acid does not require absolute purity (such as a homogeneous preparation). Instead, it represents an indication that the sequence is relatively more pure than in the natural environment (compared to the natural level this level should be at least 2- to 5-fold greater, *e.g.*, in terms of mg/mL). Individual clones isolated from a cDNA library may be purified to electrophoretic homogeneity. The claimed DNA molecules obtained from these clones could be obtained directly from total DNA or from total RNA. The cDNA clones are not naturally occurring, but rather are preferably obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The construction of a cDNA library from mRNA involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the

synthetic library by clonal selection of the cells carrying the cDNA library. Thus, the process which includes the construction of a cDNA library from mRNA and isolation of distinct cDNA clones yields an approximately 10^6-fold purification of the native message. Thus, purification of at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated.

By a " protease polypeptide" is meant 32 (preferably 40, more preferably 45, most preferably 55) or more contiguous amino acids in a polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. In certain aspects, polypeptides of 100, 200, 300, 400, 450, 500, 550, 600, 700, 800, 900 or more amino acids are preferred. The protease polypeptide can be encoded by a full-length nucleic acid sequence or any portion of the full-length nucleic acid sequence, so long as a functional activity of the polypeptide is retained. It is well known in the art that due to the degeneracy of the genetic code numerous different nucleic acid sequences can code for the same amino acid sequence. Equally, it is also well known in the art that conservative changes in amino acid can be made to arrive at a protein or polypeptide which retains the functionality of the original. Such substitutions may include the replacement of an amino acid by a residue having similar physicochemical properties, such as substituting one aliphatic residue (Ile, Val, Leu or Ala) for another, or substitution between basic residues Lys and Arg, acidic residues Glu and Asp, amide residues Gln and Asn, hydroxyl residues Ser and Tyr, or aromatic residues Phe and Tyr. Further information regarding making amino acid exchanges which have only slight, if any, effects on the overall

protein can be found in Bowie *et al.*, *Science*, 1990, 247:1306-1310, which is incorporated herein by reference in its entirety including any figures, tables, or drawings. In all cases, all permutations are intended to be covered by this disclosure.

The amino acid sequence of the protease peptide of the invention will be substantially similar to a sequence having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or the corresponding full-length amino acid sequence, or fragments thereof.

A sequence that is substantially similar to a sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70 will preferably have at least 50%, 60%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to a sequence selected from the group consisting of SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57,

SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. Preferably the protease polypeptide will have at least about 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to one of the aforementioned sequences.

By "identity" is meant a property of sequences that measures their similarity or relationship. Identity is measured by dividing the number of identical residues by the total number of residues and gaps and multiplying the product by 100. "Gaps" are spaces in an alignment that are the result of additions or deletions of amino acids. Thus, two copies of exactly the same sequence have 100% identity, but sequences that are less highly conserved, and have deletions, additions, or replacements, may have a lower degree of identity. Those skilled in the art will recognize that several computer programs are available for determining sequence identity using standard parameters, for example Gapped BLAST or PSI-BLAST (Altschul, *et al.* (1997) *Nucleic Acids Res.* 25:3389-3402), BLAST (Altschul, *et al.* (1990) *J. Mol. Biol.* 215:403-410), and Smith-Waterman (Smith, *et al.* (1981) *J. Mol. Biol.* 147:195-197). Preferably, the default settings of these programs will be employed, but those skilled in the art recognize whether these settings need to be changed and know how to make the changes.

"Similarity" is measured by dividing the number of identical residues plus the number of conservatively substituted residues (see Bowie, *et al. Science*, 1999), 247:1306-1310, which is incorporated herein by reference in its entirety, including any drawings, figures, or tables) by the total number of residues and gaps and multiplying the product by 100.

In preferred embodiments, the invention features isolated, enriched, or purified nucleic acid molecules encoding a protease polypeptide comprising a nucleotide sequence that: (a) encodes a polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID

NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70; (b) is the complement of the nucleotide sequence of (a); (c) hybridizes under highly stringent conditions to the nucleotide molecule of (a) and encodes a naturally occurring protease polypeptide.

In preferred embodiments, the invention features isolated, enriched or purified nucleic acid molecules comprising a nucleotide sequence substantially identical to a sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35. Preferably the sequence has at least 50%, 60%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to the above listed sequences.

The term "complement" refers to two nucleotides that can form multiple favorable interactions with one another. For example, adenine is complementary to thymine as they can form two hydrogen bonds. Similarly, guanine and cytosine are complementary since they can form three hydrogen bonds. A nucleotide sequence is the complement of another nucleotide sequence if all of the nucleotides of the first sequence are complementary to all of the nucleotides of the second sequence.

Various low or high stringency hybridization conditions may be used depending upon the specificity and selectivity desired. These conditions are well known to those skilled in the art. Under stringent hybridization conditions only highly complementary nucleic acid sequences hybridize. Preferably, such

conditions prevent hybridization of nucleic acids having more than 1 or 2 mismatches out of 20 contiguous nucleotides, more preferably, such conditions prevent hybridization of nucleic acids having more than 1 or 2 mismatches out of 50 contiguous nucleotides, most preferably, such conditions prevent hybridization of nucleic acids having more than 1 or 2 mismatches out of 100 contiguous nucleotides. In some instances, the conditions may prevent hybridization of nucleic acids having more than 5 mismatches in the full-length sequence.

By stringent hybridization assay conditions is meant hybridization assay conditions at least as stringent as the following: hybridization in 50% formamide, 5X SSC, 50 mM NaH_2PO_4, pH 6.8, 0.5% SDS, 0.1 mg/mL sonicated salmon sperm DNA, and 5X Denhardt's solution at 42 °C overnight; washing with 2X SSC, 0.1% SDS at 45 °C; and washing with 0.2X SSC, 0.1% SDS at 45 °C. Under some of the most stringent hybridization assay conditions, the second wash can be done with 0.1X SSC at a temperature up to 70 °C (Berger *et al.* (1987) Guide to Molecular Cloning Techniques pg 421, hereby incorporated by reference herein in its entirety including any figures, tables, or drawings.). However, other applications may require the use of conditions falling between these sets of conditions. Methods of determining the conditions required to achieve desired hybridizations are well known to those with ordinary skill in the art, and are based on several factors, including but not limited to, the sequences to be hybridized and the samples to be tested. Washing conditions of lower stringency frequently utilize a lower temperature during the washing steps, such as 65 °C, 60 °C, 55 °C, 50 °C, or 42 °C.

The term "activity" means that the polypeptide hydrolyzes peptide bonds.

The term "catalytic activity", as used herein, defines the rate at which a protease catalytic domain cleaves a substrate. Catalytic activity can be measured, for example, by determining the amount of a substrate cleaved as a function of time. Catalytic activity can be measured by methods of the invention by holding time constant and determining the concentration of a cleaved substrate after a fixed period of time. Cleavage of a substrate occurs at the active site of the protease. The

active site is normally a cavity in which the substrate binds to the protease and is cleaved.

The term "substrate" as used herein refers to a polypeptide or protein which is cleaved by a protease of the invention. The term "cleaved" refers to the severing of a covalent bond between amino acid residues of the backbone of the polypeptide or protein.

The term "insert" as used herein refers to a portion of a protease that is absent from a close homolog. Inserts may or may not -be the product alternative splicing of exons. Inserts can be identified by using a Smith-Waterman sequence alignment of the protein sequence against the non-redundant protein database, or by means of a multiple sequence alignment of homologous sequences using the DNAStar program Megalign (Preferably, the default settings of this program will be used, but those skilled in the art will recognize whether these settings need to be changed and know how to make the changes.). Inserts may play a functional role by presenting a new interface for protein-protein interactions, or by interfering with such interactions.

In other preferred embodiments, the invention features isolated, enriched, or purified nucleic acid molecules encoding protease polypeptides, further comprising a vector or promoter effective to initiate transcription in a host cell. The invention also features recombinant nucleic acid, preferably in a cell or an organism. The recombinant nucleic acid may contain a sequence selected from the group consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35, or a functional derivative thereof and a vector or a promoter effective to initiate transcription in a host cell. The recombinant nucleic acid can

alternatively contain a transcriptional initiation region functional in a cell, a sequence complementary to an RNA sequence encoding a protease polypeptide and a transcriptional termination region functional in a cell. Specific vectors and host cell combinations are discussed herein.

The term "vector" relates to a single or double-stranded circular nucleic acid molecule that can be transfected into cells and replicated within or independently of a cell genome. A circular double-stranded nucleic acid molecule can be cut and thereby linearized upon treatment with restriction enzymes. An assortment of nucleic acid vectors, restriction enzymes, and the knowledge of the nucleotide sequences cut by restriction enzymes are readily available to those skilled in the art. A nucleic acid molecule encoding a protease can be inserted into a vector by cutting the vector with restriction enzymes and ligating the two pieces together.

The term "transfecting" defines a number of methods to insert a nucleic acid vector or other nucleic acid molecules into a cellular organism. These methods involve a variety of techniques, such as treating the cells with high concentrations of salt, an electric field, detergent, or DMSO to render the outer membrane or wall of the cells permeable to nucleic acid molecules of interest or use of various viral transduction strategies.

The term "promoter" as used herein, refers to nucleic acid sequence needed for gene sequence expression. Promoter regions vary from organism to organism, but are well known to persons skilled in the art for different organisms. For example, in prokaryotes, the promoter region contains both the promoter (which directs the initiation of RNA transcription) as well as the DNA sequences which, when transcribed into RNA, will signal synthesis initiation. Such regions will normally include those 5'-non-coding sequences involved with initiation of transcription and translation, such as the TATA box, capping sequence, CAAT sequence, and the like.

In preferred embodiments, the isolated nucleic acid comprises, consists essentially of, or consists of a nucleic acid sequence selected from the group consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ

ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35 which encodes an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, a functional derivative thereof, or at least 35, 40, 45, 50, 60, 75, 100, 200, or 300 contiguous amino acids selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. The nucleic acid may be isolated from a natural source by cDNA cloning or by subtractive hybridization. The natural source may be mammalian, preferably human, blood, semen, or tissue, and the nucleic acid may be synthesized by the triester method or by using an automated DNA synthesizer.

The term "mammal" refers preferably to such organisms as mice, rats, rabbits, guinea pigs, sheep, and goats, more preferably to cats, dogs, monkeys, and apes, and most preferably to humans.

In yet other preferred embodiments, the nucleic acid is a conserved or unique region, for example those useful for: the design of hybridization probes to facilitate identification and cloning of additional polypeptides, the design of PCR probes to facilitate cloning of additional polypeptides, obtaining antibodies to polypeptide regions, and designing antisense oligonucleotides.

By "conserved nucleic acid regions", are meant regions present on two or more nucleic acids encoding a protease polypeptide, to which a particular nucleic acid sequence can hybridize under lower stringency conditions. Examples of lower stringency conditions suitable for screening for nucleic acid encoding protease polypeptides are provided in Wahl *et al. Meth. Enzym.* 152:399-407 (1987) and in Wahl *et al. Meth. Enzym.* 152:415-423 (1987), which are hereby incorporated by reference herein in its entirety, including any drawings, figures, or tables. Preferably, conserved regions differ by no more than 5 out of 20 nucleotides, even more preferably 2 out of 20 nucleotides or most preferably 1 out of 20 nucleotides.

By "unique nucleic acid region" is meant a sequence present in a nucleic acid coding for a protease polypeptide that is not present in a sequence coding for any other naturally occurring polypeptide. Such regions preferably encode 32 (preferably 40, more preferably 45, most preferably 55) or more contiguous amino acids set forth in a full-length amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70 in a sample. The nucleic acid probe contains a

nucleotide base sequence that will hybridize to the sequence selected from the group consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35, or a functional derivative thereof.

In preferred embodiments, the nucleic acid probe hybridizes to nucleic acid encoding at least 12, 32, 75, 90, 105, 120, 150, 200, 250, 300 or 350 contiguous amino acids of a full-length sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or a functional derivative thereof.

Methods for using the probes include detecting the presence or amount of protease RNA in a sample by contacting the sample with a nucleic acid probe under conditions such that hybridization occurs and detecting the presence or amount of the probe bound to protease RNA. The nucleic acid duplex formed between the probe and a nucleic acid sequence coding for a protease polypeptide may be used in the identification of the sequence of the nucleic acid detected (Nelson *et al.*, in Nonisotopic DNA Probe Techniques, Academic Press, San Diego, Kricka, ed., p. 275, 1992, hereby incorporated by reference herein in its entirety, including any drawings, figures, or tables). Kits for performing such methods may be constructed to include a container means having disposed therein a nucleic acid probe.

Methods for using the probes also include using these probes to find the full-length clone of each of the predicted proteases by techniques known to one skilled in the art. These clones will be useful for screening for small molecule compounds that inhibit the catalytic activity of the encoded protease with potential utility in treating cancers, immune-related diseases and disorders, cardiovascular disease, brain or neuronal-associated diseases, and metabolic disorders. More specifically disorders including cancers of tissues, blood, or hematopoietic origin, particularly those involving breast, colon, lung, prostate, cervical, brain, ovarian, bladder, or kidney; central or peripheral nervous system diseases and conditions including migraine, pain, sexual dysfunction, mood disorders, attention disorders, cognition disorders, hypotension, and hypertension; psychotic and neurological disorders, including anxiety, schizophrenia, manic depression, delirium, dementia, severe mental retardation and dyskinesias, such as Huntington's disease or Tourette's Syndrome; neurodegenerative diseases including Alzheimer's, Parkinson's, multiple sclerosis, and amyotrophic lateral sclerosis; viral or non-viral infections caused by HIV-1, HIV-2 or other viral- or prion-agents or fungal- or bacterial- organisms; metabolic disorders including Diabetes and obesity and their related syndromes, among others; cardiovascular disorders including reperfusion restenosis, coronary thrombosis, clotting disorders, unregulated cell growth disorders, atherosclerosis; ocular disease including glaucoma, retinopathy, and macular degeneration; inflammatory disorders including rheumatoid arthritis, chronic inflammatory bowel disease, chronic inflammatory pelvic disease, multiple sclerosis, asthma, osteoarthritis, psoriasis, atherosclerosis, rhinitis, autoimmunity, and organ transplant rejection.

In another aspect, the invention describes a recombinant cell or tissue comprising a nucleic acid molecule encoding a protease polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID

NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. In such cells, the nucleic acid may be under the control of the genomic regulatory elements, or may be under the control of exogenous regulatory elements including an exogenous promoter. By "exogenous" it is meant a promoter that is not normally coupled *in vivo* transcriptionally to the coding sequence for the protease polypeptides.

The polypeptide is preferably a fragment of the protein encoded by a full-length amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. By "fragment," is meant an amino acid sequence present in a protease polypeptide. Preferably, such a sequence comprises at least 32, 45, 50, 60, 100, 200, or 300 contiguous amino acids of a full-length sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

In another aspect, the invention features an isolated, enriched, or purified protease polypeptide having the amino acid sequence selected from the group

consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

By "isolated" in reference to a polypeptide is meant a polymer of 6 (preferably 12, more preferably 18, most preferably 25, 32, 40, or 50) or more amino acids conjugated to each other, including polypeptides that are isolated from a natural source or that are synthesized. In certain aspects longer polypeptides are preferred, such as those with 100, 200, 300, 400, 450, 500, 550, 600, 700, 800, 900 or more contiguous amino acids of a full-length sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, and/or those polypeptides having at least 50%, 60%, 75%, 80%, 85%, 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identity to a sequence selected from the group consisting of SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID

NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

The isolated polypeptides of the present invention are unique in the sense that they are not found in a pure or separated state in nature. Use of the term "isolated" indicates that a naturally occurring sequence has been removed from its normal cellular environment. Thus, the sequence may be in a cell-free solution or placed in a different cellular environment. The term does not imply that the sequence is the only amino acid chain present, but that it is essentially free (at least about 90% pure, more preferably at least about 95% pure or more) of non-amino acid-based material naturally associated with it.

By the use of the term "enriched" in reference to a polypeptide is meant that the specific amino acid sequence constitutes a significantly higher fraction (2- to 5-fold) of the total amino acid sequences present in the cells or solution of interest than in normal or diseased cells or in the cells from which the sequence was taken. This could be caused by a person by preferential reduction in the amount of other amino acid sequences present, or by a preferential increase in the amount of the specific amino acid sequence of interest, or by a combination of the two. However, it should be noted that enriched does not imply that there are no other amino acid sequences present, just that the relative amount of the sequence of interest has been significantly increased. The term significant here is used to indicate that the level of increase is useful to the person making such an increase, and generally means an increase relative to other amino acid sequences of about at least 2-fold, more preferably at least 5- to 10-fold or even more. The term also does not imply that there is no amino acid sequence from other sources. The other source of amino acid sequences may, for example, comprise amino acid sequence encoded by a yeast or bacterial genome, or a cloning vector such as pUC19. The term is meant to cover only those situations in which man has intervened to increase the proportion of the desired amino acid sequence.

It is also advantageous for some purposes that an amino acid sequence be in purified form. The term "purified" in reference to a polypeptide does not require

absolute purity (such as a homogeneous preparation); instead, it represents an indication that the sequence is relatively purer than in the natural environment. Compared to the natural level this level should be at least 2-to 5-fold greater (*e.g.*, in terms of mg/mL). Purification of at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly contemplated. The substance is preferably free of contamination at a functionally significant level, for example 90%, 95%, or 99% pure.

In preferred embodiments, the protease polypeptide is a fragment of the protein encoded by a full-length amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. Preferably, the protease polypeptide contains at least 32, 45, 50, 60, 100, 200, or 300 contiguous amino acids of a full-length sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or a functional derivative thereof.

In preferred embodiments, the protease polypeptide comprises an amino acid sequence having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ

ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

The polypeptide can be isolated from a natural source by methods well-known in the art. The natural source may be mammalian, preferably human, blood, semen, or tissue, and the polypeptide may be synthesized using an automated polypeptide synthesizer.

In some embodiments the invention includes a recombinant protease polypeptide having (a) an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. By "recombinant protease polypeptide" is meant a polypeptide produced by recombinant DNA techniques such that it is distinct from a naturally occurring polypeptide either in its location (*e.g.*, present in a different cell or tissue than found in nature), purity or structure. Generally, such a recombinant polypeptide will be present in a cell in an amount different from that normally observed in nature.

The polypeptides to be expressed in host cells may also be fusion proteins which include regions from heterologous proteins. Such regions may be included to allow, *e.g.*, secretion, improved stability, or facilitated purification of the polypeptide. For example, a sequence encoding an appropriate signal peptide can be

incorporated into expression vectors. A DNA sequence for a signal peptide (secretory leader) may be fused in-frame to the polynucleotide sequence so that the polypeptide is translated as a fusion protein comprising the signal peptide. A signal peptide that is functional in the intended host cell promotes extracellular secretion of the polypeptide. Preferably, the signal sequence will be cleaved from the polypeptide upon secretion of the polypeptide from the cell. Thus, preferred fusion proteins can be produced in which the N-terminus of a protease polypeptide is fused to a carrier peptide.

In one embodiment, the polypeptide comprises a fusion protein which includes a heterologous region used to facilitate purification of the polypeptide. Many of the available peptides used for such a function allow selective binding of the fusion protein to a binding partner. A preferred binding partner includes one or more of the IgG binding domains of protein A are easily purified to homogeneity by affinity chromatography on, for example, IgG-coupled Sepharose. Alternatively, many vectors have the advantage of carrying a stretch of histidine residues that can be expressed at the N-terminal or C-terminal end of the target protein, and thus the protein of interest can be recovered by metal chelation chromatography. A nucleotide sequence encoding a recognition site for a proteolytic enzyme such as enterokinase, factor X procollagenase or thrombine may immediately precede the sequence for a protease polypeptide to permit cleavage of the fusion protein to obtain the mature protease polypeptide. Additional examples of fusion-protein binding partners include, but are not limited to, the yeast I-factor, the honeybee melatin leader in sf9 insect cells, 6-His tag, thioredoxin tag, hemaglutinin tag, GST tag, and OmpA signal sequence tag. As will be understood by one of skill in the art, the binding partner which recognizes and binds to the peptide may be any ion, molecule or compound including metal ions (*e.g.*, metal affinity columns), antibodies, or fragments thereof, and any protein or peptide which binds the peptide, such as the FLAG tag.

Antibodies

In another aspect, the invention features an antibody (*e.g.*, a monoclonal or polyclonal antibody) having specific binding affinity to a protease polypeptide or a protease polypeptide domain or fragment where the polypeptide is selected from the group having a sequence at least about 90% identical to an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. By "specific binding affinity" is meant that the antibody binds to the target protease polypeptide with greater affinity than it binds to other polypeptides under specified conditions. Antibodies or antibody fragments are polypeptides that contain regions that can bind other polypeptides. The term "specific binding affinity" describes an antibody that binds to a protease polypeptide with greater affinity than it binds to other polypeptides under specified conditions. Antibodies can be used to identify an endogenous source of protease polypeptides, to monitor cell cycle regulation, and for immuno-localization of protease polypeptides within the cell.

The term "polyclonal" refers to antibodies that are heterogenous populations of antibody molecules derived from the sera of animals immunized with an antigen or an antigenic functional derivative thereof. For the production of polyclonal antibodies, various host animals may be immunized by injection with the antigen. Various adjuvants may be used to increase the immunological response, depending on the host species.

"Monoclonal antibodies" are substantially homogenous populations of antibodies to a particular antigen. They may be obtained by any technique which provides for the production of antibody molecules by continuous cell lines in

culture. Monoclonal antibodies may be obtained by methods known to those skilled in the art (Kohler *et al.*, *Nature*, 1975, 256:495-497, and U.S. Patent No. 4,376,110, both of which are hereby incorporated by reference herein in their entirety including any figures, tables, or drawings).

An antibody of the present invention includes "humanized" monoclonal and polyclonal antibodies. Humanized antibodies are recombinant proteins in which non-human (typically murine) complementarity determining regions of an antibody have been transferred from heavy and light variable chains of the non-human (*e.g.* murine) immunoglobulin into a human variable domain, followed by the replacement of some human residues in the framework regions of their murine counterparts. Humanized antibodies in accordance with this invention are suitable for use in therapeutic methods. General techniques for cloning murine immunoglobulin variable domains are described, for example, by the publication of Orlandi *et al., Proc. Nat'l Acad. Sci. USA 86*: 3833 (1989). Techniques for producing humanized monoclonal antibodies are described, for example, by Jones *et al., Nature 321*:522 (1986), Riechmann *et al., Nature 332*:323 (1988), Verhoeyen *et al., Science 239*:1534 (1988), Carter *et al., Proc. Nat'l Acad. Sci. USA 89*:4285 (1992), Sandhu, *Crit. Rev. Biotech. 12*:437 (1992), and Singer *et al., J. Immun. 150*:2844 (1993).

The term "antibody fragment" refers to a portion of an antibody, often the hypervariable region and portions of the surrounding heavy and light chains, that displays specific binding affinity for a particular molecule. A hypervariable region is a portion of an antibody that physically binds to the polypeptide target.

An antibody fragment of the present invention includes a "single-chain antibody," a phrase used in this description to denote a linear polypeptide that binds antigen with specificity and that comprises variable or hypervariable regions from the heavy and light chain chains of an antibody. Such single chain antibodies can be produced by conventional methodology. The Vh and Vl regions of the Fv fragment can be covalently joined and stabilized by the insertion of a disulfide bond. See Glockshuber, *et al.*, *Biochemistry* 1362 (1990). Alternatively, the Vh and Vl regions

can be joined by the insertion of a peptide linker. A gene encoding the Vh, Vl and peptide linker sequences can be constructed and expressed using a recombinant expression vector. See Colcher, *et al.*, *J. Nat'l Cancer Inst.* 82: 1191 (1990). Amino acid sequences comprising hypervariable regions from the Vh and Vl antibody chains can also be constructed using disulfide bonds or peptide linkers.

Antibodies or antibody fragments having specific binding affinity to a protease polypeptide of the invention may be used in methods for detecting the presence and/or amount of protease polypeptide in a sample by probing the sample with the antibody under conditions suitable for protease-antibody immunocomplex formation and detecting the presence and/or amount of the antibody conjugated to the protease polypeptide. Diagnostic kits for performing such methods may be constructed to include antibodies or antibody fragments specific for the protease as well as a conjugate of a binding partner of the antibodies or the antibodies themselves.

An antibody or antibody fragment with specific binding affinity to a protease polypeptide of the invention can be isolated, enriched, or purified from a prokaryotic or eukaryotic organism. Routine methods known to those skilled in the art enable production of antibodies or antibody fragments, in both prokaryotic and eukaryotic organisms. Purification, enrichment, and isolation of antibodies, which are polypeptide molecules, are described above.

Antibodies having specific binding affinity to a protease polypeptide of the invention may be used in methods for detecting the presence and/or amount of protease polypeptide in a sample by contacting the sample with the antibody under conditions such that an immunocomplex forms and detecting the presence and/or amount of the antibody conjugated to the protease polypeptide. Diagnostic kits for performing such methods may be constructed to include a first container containing the antibody and a second container having a conjugate of a binding partner of the antibody and a label, such as, for example, a radioisotope. The diagnostic kit may also include notification of an FDA approved use and instructions therefor.

In another aspect, the invention features a hybridoma which produces an antibody having specific binding affinity to a protease polypeptide or a protease polypeptide domain, where the polypeptide is selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. By "hybridoma" is meant an immortalized cell line that is capable of secreting an antibody, for example an antibody to a protease of the invention. In preferred embodiments, the antibody to the protease comprises a sequence of amino acids that is able to specifically bind a protease polypeptide of the invention.

In another aspect, the present invention is also directed to kits comprising antibodies that bind to a polypeptide encoded by any of the nucleic acid molecules described above, and a negative control antibody.

The term "negative control antibody" refers to an antibody derived from similar source as the antibody having specific binding affinity, but where it displays no binding affinity to a polypeptide of the invention.

In another aspect, the invention features a protease polypeptide binding agent able to bind to a protease polypeptide selected from the group having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66,

SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. The binding agent is preferably a purified antibody that recognizes an epitope present on a protease polypeptide of the invention. Other binding agents include molecules that bind to protease polypeptides and analogous molecules that bind to a protease polypeptide. Such binding agents may be identified by using assays that measure protease binding partner activity, or they may be identified using assays that measure protease activity, such as the release of a fluorogenic or radioactive marker attached to a substrate molecule.

Screening Methods to Detect Protease Polypeptides

The invention also features a method for screening for human cells containing a protease polypeptide of the invention or an equivalent sequence. The method involves identifying the novel polypeptide in human cells using techniques that are routine and standard in the art, such as those described herein for identifying the proteases of the invention (*e.g.*, cloning, Southern or Northern blot analysis, *in situ* hybridization, PCR amplification, etc.).

Screening Methods to Identify Substances that Modulate Protease Activity

In another aspect, the invention features methods for identifying a substance that modulates protease activity comprising the steps of: (a) contacting a protease polypeptide comprising an amino acid substantially identical to a sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70 with a test substance; (b) measuring the activity of said polypeptide; and (c) determining whether said substance modulates the activity of said polypeptide. More preferably the sequence is at least

about 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98% or 99% identical to the listed sequences.

The term "modulates" refers to the ability of a compound to alter the function of a protease of the invention. A modulator preferably activates or inhibits the activity of a protease of the invention depending on the concentration of the compound exposed to the protease.

The term "modulates" also refers to altering the function of proteases of the invention by increasing or decreasing the probability that a complex forms between the protease and a natural binding partner. A modulator preferably increases the probability that such a complex forms between the protease and the natural binding partner, more preferably increases or decreases the probability that a complex forms between the protease and the natural binding partner depending on the concentration of the compound exposed to the protease, and most preferably decreases the probability that a complex forms between the protease and the natural binding partner.

The term "activates" refers to increasing the cellular activity of the protease. The term "inhibits" refers to decreasing the cellular activity of the protease.

The term "complex" refers to an assembly of at least two molecules bound to one another. Signal transduction complexes often contain at least two protein molecules bound to one another. For instance, a protein tyrosine receptor protein kinase, GRB2, SOS, RAF, and RAS assemble to form a signal transduction complex in response to a mitogenic ligand. Similarly, the proteases involved in blood coagulation and their cofactors are known to form macromolecular complexes on cellular membranes. Additionally, proteases involved in modification of the extracellular matrix are known to form complexes with their inhibitors and also with components of the extracellular matrix.

The term "natural binding partner" refers to polypeptides, lipids, small molecules, or nucleic acids that bind to proteases in cells. A change in the interaction between a protease and a natural binding partner can manifest itself as an

increased or decreased probability that the interaction forms, or an increased or decreased concentration of protease/natural binding partner complex.

The term "contacting" as used herein refers to mixing a solution comprising the test compound with a liquid medium bathing the cells of the methods. The solution comprising the compound may also comprise another component, such as dimethyl sulfoxide (DMSO), which facilitates the uptake of the test compound or compounds into the cells of the methods. The solution comprising the test compound may be added to the medium bathing the cells by utilizing a delivery apparatus, such as a pipette-based device or syringe-based device.

In another aspect, the invention features methods for identifying a substance that modulates protease activity in a cell comprising the steps of: (a) expressing a protease polypeptide in a cell, wherein said polypeptide is selected from the group having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70; (b) adding a test substance to said cell; and (c) monitoring a change in cell phenotype or the interaction between said polypeptide and a natural binding partner.

The term "expressing" as used herein refers to the production of proteases of the invention from a nucleic acid vector containing protease genes within a cell. The nucleic acid vector is transfected into cells using well known techniques in the art as described herein.

Another aspect of the instant invention is directed to methods of identifying compounds that bind to protease polypeptides of the present invention, comprising contacting the protease polypeptides with a compound, and determining whether the

compound binds the protease polypeptides. Binding can be determined by binding assays which are well known to the skilled artisan, including, but not limited to, gel-shift assays, Western blots, radiolabeled competition assay, phage-based expression cloning, co-fractionation by chromatography, co-precipitation, cross linking, interaction trap/two-hybrid analysis, southwestern analysis, ELISA, and the like, which are described in, for example, *Current Protocols in Molecular Biology*, 1999, John Wiley & Sons, NY, which is incorporated herein by reference in its entirety. The compounds to be screened include, but are not limited to, compounds of extracellular, intracellular, biological or chemical origin.

The methods of the invention also embrace compounds that are attached to a label, such as a radiolabel (*e.g.*, ^{125}I, ^{35}S, ^{32}P, ^{33}P, ^{3}H), a fluorescence label, a chemiluminescent label, an enzymic label and an immunogenic label. The protease polypeptides employed in such a test may either be free in solution, attached to a solid support, borne on a cell surface, located intracellularly or associated with a portion of a cell. One skilled in the art can, for example, measure the formation of complexes between a protease polypeptide and the compound being tested. Alternatively, one skilled in the art can examine the diminution in complex formation between a protease polypeptide and its substrate caused by the compound being tested.

Other assays can be used to examine enzymatic activity including, but not limited to, photometric, radiometric, HPLC, electrochemical, and the like, which are described in, for example, Enzyme Assays: A Practical Approach, eds. R. Eisenthal and M. J. Danson, 1992, Oxford University Press, which is incorporated herein by reference in its entirety.

Another aspect of the present invention is directed to methods of identifying compounds which modulate (*i.e.*, increase or decrease) activity of a protease polypeptide comprising contacting the protease polypeptide with a compound, and determining whether the compound modifies activity of the protease polypeptide. These compounds are also referred to as "modulators of proteases." The activity in the presence of the test compound is measured to the activity in the absence of the

test compound. Where the activity of a sample containing the test compound is higher than the activity in a sample lacking the test compound, the compound will have increased the activity. Similarly, where the activity of a sample containing the test compound is lower than the activity in the sample lacking the test compound, the compound will have inhibited the activity.

The present invention is particularly useful for screening compounds by using a protease polypeptide in any of a variety of drug screening techniques. The compounds to be screened include, but are not limited to, extracellular, intracellular, biological or chemical origin. The protease polypeptide employed in such a test may be in any form, preferably, free in solution, attached to a solid support, borne on a cell surface or located intracellularly. One skilled in the art can measure the change in rate that a protease of the invention cleaves a substrate polypeptide. One skilled in the art can also, for example, measure the formation of complexes between a protease polypeptide and the compound being tested. Alternatively, one skilled in the art can examine the diminution in complex formation between a protease polypeptide and its substrate caused by the compound being tested.

The activity of protease polypeptides of the invention can be determined by, for example, examining the ability to bind or be activated by chemically synthesised peptide ligands. Alternatively, the activity of the protease polypeptides can be assayed by examining their ability to bind metal ions such as calcium, hormones, chemokines, neuropeptides, neurotransmitters, nucleotides, lipids, odorants, and photons. Thus, modulators of the protease polypeptide's activity may alter a protease function, such as a binding property of a protease or an activity such as cleaving protein substrates or polypeptide substrates, or membrane localization.

In various embodiments of the method, the assay may take the form of a yeast growth assay, an Aequorin assay, a Luciferase assay, a mitogenesis assay, a MAP Kinase activity assay, as well as other binding or function-based assays of protease activity that are generally known in the art. In several of these embodiments, the invention includes any of the serine proteases, cysteine proteases, aspartyl proteases, metalloproteases, threonine proteases, and other proteases.

Biological activities of proteases according to the invention include, but are not limited to, the binding of a natural or a synthetic ligand, as well as any one of the functional activities of proteases known in the art. Non-limiting examples of protease activities include cleavage of polypeptide chains, processing the pro-form of a polypeptide chain to the active product, transmembrane signaling of various forms, and/or the modification of the extraceullar matrix.

The modulators of the invention exhibit a variety of chemical structures, which can be generally grouped into mimetics of natural protease ligands, and peptide and non-peptide allosteric effectors of proteases. The invention does not restrict the sources for suitable modulators, which may be obtained from natural sources such as plant, animal or mineral extracts, or non-natural sources such as small molecule libraries, including the products of combinatorial chemical approaches to library construction, and peptide libraries.

The use of cDNAs encoding proteins in drug discovery programs is well-known; assays capable of testing thousands of unknown compounds per day in high-throughput screens (HTSs) are thoroughly documented. The literature is replete with examples of the use of radiolabelled ligands in HTS binding assays for drug discovery (*see*, Williams, *Medicinal Research Reviews*, 1991, 11:147-184.; Sweetnam, *et al.*, *J. Natural Products*, 1993, 56:441-455 for review). Recombinant proteins are preferred for binding assay HTS because they allow for better specificity (higher relative purity), provide the ability to generate large amounts of receptor material, and can be used in a broad variety of formats (see Hodgson, *Bio/Technology*, 1992, 10:973-980 which is incorporated herein by reference in its entirety). A variety of heterologous systems is available for functional expression of recombinant proteins that are well known to those skilled in the art. Such systems include bacteria (Strosberg, *et al.*, *Trends in Pharmacological Sciences*, 1992, 13:95-98), yeast (Pausch, *Trends in Biotechnology*, 1997, 15:487-494), several kinds of insect cells (Vanden Broeck, *Int. Rev. Cytology*, 1996, 164:189-268), amphibian cells (Jayawickreme *et al.*, *Current Opinion in Biotechnology*, 1997, 8:629-634) and several mammalian cell lines (CHO, HEK293, COS, etc.; *see*, Gerhardt, *et al.*, *Eur.*

J. Pharmacology, 1997, 334:1-23). These examples do not preclude the use of other possible cell expression systems, including cell lines obtained from nematodes (PCT application WO 98/37177).

An expressed protease can be used for HTS binding assays in conjunction with its defined ligand, in this case the corresponding peptide that activates it. The identified peptide is labeled with a suitable radioisotope, including, but not limited to, ^{125}I, ^{3}H, ^{35}S or ^{32}P, by methods that are well known to those skilled in the art. Alternatively, the peptides may be labeled by well-known methods with a suitable fluorescent derivative (Baindur, *et al.*, *Drug Dev. Res.*, 1994, 33:373-398; Rogers, *Drug Discovery Today*, 1997, 2:156-160). Radioactive ligand specifically bound to the receptor in membrane preparations made from the cell line expressing the recombinant protein can be detected in HTS assays in one of several standard ways, including filtration of the receptor-ligand complex to separate bound ligand from unbound ligand (Williams, *Med. Res. Rev.*, 1991, 11:147-184.; Sweetnam, *et al.*, *J. Natural Products*, 1993, 56:441-455). Alternative methods include a scintillation proximity assay (SPA) or a FlashPlate format in which such separation is unnecessary (Nakayama, *Cur. Opinion Drug Disc. Dev.*, 1998, 1:85-91 Bossé, *et al.*, *J. Biomolecular Screening*, 1998, 3:285-292.). Binding of fluorescent ligands can be detected in various ways, including fluorescence energy transfer (FRET), direct spectrophotofluorometric analysis of bound ligand, or fluorescence polarization (Rogers, *Drug Discovery Today*, 1997, 2:156-160; Hill, *Cur. Opinion Drug Disc. Dev.*, 1998, 1:92-97).

The proteases and natural binding partners required for functional expression of heterologous protease polypeptides can be native constituents of the host cell or can be introduced through well-known recombinant technology. The protease polypeptides can be intact or chimeric. The protease activation may result in the stimulation or inhibition of other native proteins, events that can be linked to a measurable response.

Examples of such biological responses include, but are not limited to, the following: the ability to survive in the absence of a limiting nutrient in specifically

engineered yeast cells (Pausch, *Trends in Biotechnology*, 1997, 15:487-494); changes in intracellular Ca^{2+} concentration as measured by fluorescent dyes (Murphy, *et al.*, *Cur. Opinion Drug Disc. Dev.*, 1998, 1:192-199). Fluorescence changes can also be used to monitor ligand-induced changes in membrane potential or intracellular pH; an automated system suitable for HTS has been described for these purposes (Schroeder, *et al.*, *J. Biomolecular Screening*, 1996, 1:75-80). Assays are also available for the measurement of common second but these are not generally preferred for HTS.

The invention contemplates a multitude of assays to screen and identify inhibitors of ligand binding to protease polypeptides or of substrate cleavage by protease polypeptides. In one example, the protease polypeptide is immobilized and interaction with a binding partner or substrate is assessed in the presence and absence of a candidate modulator such as an inhibitor compound. In another example, interaction between the protease polypeptide and its binding partner or a substrate is assessed in a solution assay, both in the presence and absence of a candidate inhibitor compound. In either assay, an inhibitor is identified as a compound that decreases binding between the protease polypeptide and its natural binding partner or the activity of a protease polypeptide in cleaving a substrate molecule. Another contemplated assay involves a variation of the di-hybrid assay wherein an inhibitor of protein/protein interactions is identified by detection of a positive signal in a transformed or transfected host cell, as described in PCT publication number WO 95/20652, published August 3, 1995 and is included by reference herein including any figures, tables, or drawings.

Candidate modulators contemplated by the invention include compounds selected from libraries of either potential activators or potential inhibitors. There are a number of different libraries used for the identification of small molecule modulators, including: (1) chemical libraries, (2) natural product libraries, and (3) combinatorial libraries comprised of random peptides, oligonucleotides or organic molecules. Chemical libraries consist of random chemical structures, some of which are analogs of known compounds or analogs of compounds that have been identified

as "hits" or "leads" in other drug discovery screens, while others are derived from natural products, and still others arise from non-directed synthetic organic chemistry. Natural product libraries are collections of microorganisms, animals, plants, or marine organisms which are used to create mixtures for screening by: (1) fermentation and extraction of broths from soil, plant or marine microorganisms or (2) extraction of plants or marine organisms. Natural product libraries include polyketides, non-ribosomal peptides, and variants (non-naturally occurring) thereof. For a review, *see, Science* 282:63-68 (1998). Combinatorial libraries are composed of large numbers of peptides, oligonucleotides, or organic compounds as a mixture. These libraries are relatively easy to prepare by traditional automated synthesis methods, PCR, cloning, or proprietary synthetic methods. Of particular interest are non-peptide combinatorial libraries. Still other libraries of interest include peptide, protein, peptidomimetic, multiparallel synthetic collection, recombinatorial, and polypeptide libraries. For a review of combinatorial chemistry and libraries created therefrom, *see,* Myers, *Curr. Opin. Biotechnol.* 8:701-707 (1997). Identification of modulators through use of the various libraries described herein permits modification of the candidate "hit" (or "lead") to optimize the capacity of the "hit" to modulate activity.

Still other candidate inhibitors contemplated by the invention can be designed and include soluble forms of binding partners, as well as such binding partners as chimeric, or fusion, proteins. A "binding partner" as used herein broadly encompasses both natural binding partners as described above as well as chimeric polypeptides, peptide modulators other than natural ligands, antibodies, antibody fragments, and modified compounds comprising antibody domains that are immunospecific for the expression product of the identified protease gene.

Other assays may be used to identify specific peptide ligands of a protease polypeptide, including assays that identify ligands of the target protein through measuring direct binding of test ligands to the target protein, as well as assays that identify ligands of target proteins through affinity ultrafiltration with ion spray mass spectroscopy/HPLC methods or other physical and analytical methods.

Alternatively, such binding interactions are evaluated indirectly using the yeast two-hybrid system described in Fields *et al.*, *Nature*, 340:245-246 (1989), and Fields *et al.*, *Trends in Genetics*, 10:286-292 (1994), both of which are incorporated herein by reference. The two-hybrid system is a genetic assay for detecting interactions between two proteins or polypeptides. It can be used to identify proteins that bind to a known protein of interest, or to delineate domains or residues critical for an interaction. Variations on this methodology have been developed to clone genes that encode DNA binding proteins, to identify peptides that bind to a protein, and to screen for drugs. The two-hybrid system exploits the ability of a pair of interacting proteins to bring a transcription activation domain into close proximity with a DNA binding domain that binds to an upstream activation sequence (UAS) of a reporter gene, and is generally performed in yeast. The assay requires the construction of two hybrid genes encoding (1) a DNA-binding domain that is fused to a first protein and (2) an activation domain fused to a second protein. The DNA-binding domain targets the first hybrid protein to the UAS of the reporter gene; however, because most proteins lack an activation domain, this DNA-binding hybrid protein does not activate transcription of the reporter gene. The second hybrid protein, which contains the activation domain, cannot by itself activate expression of the reporter gene because it does not bind the UAS. However, when both hybrid proteins are present, the noncovalent interaction of the first and second proteins tethers the activation domain to the UAS, activating transcription of the reporter gene. For example, when the first protein is a protease gene product, or fragment thereof, that is known to interact with another protein or nucleic acid, this assay can be used to detect agents that interfere with the binding interaction. Expression of the reporter gene is monitored as different test agents are added to the system. The presence of an inhibitory agent results in lack of a reporter signal.

When the function of the protease polypeptide gene product is unknown and no ligands are known to bind the gene product, the yeast two-hybrid assay can also be used to identify proteins that bind to the gene product. In an assay to identify proteins that bind to a protease polypeptide, or fragment thereof, a fusion

polynucleotide encoding both a protease polypeptide (or fragment) and a UAS binding domain (*i.e.*, a first protein) may be used. In addition, a large number of hybrid genes each encoding a different second protein fused to an activation domain are produced and screened in the assay. Typically, the second protein is encoded by one or more members of a total cDNA or genomic DNA fusion library, with each second protein coding region being fused to the activation domain. This system is applicable to a wide variety of proteins, and it is not even necessary to know the identity or function of the second binding protein. The system is highly sensitive and can detect interactions not revealed by other methods; even transient interactions may trigger transcription to produce a stable mRNA that can be repeatedly translated to yield the reporter protein.

Other assays may be used to search for agents that bind to the target protein. One such screening method to identify direct binding of test ligands to a target protein is described in U.S. Patent No. 5,585,277, incorporated herein by reference. This method relies on the principle that proteins generally exist as a mixture of folded and unfolded states, and continually alternate between the two states. When a test ligand binds to the folded form of a target protein (*i.e.*, when the test ligand is a ligand of the target protein), the target protein molecule bound by the ligand remains in its folded state. Thus, the folded target protein is present to a greater extent in the presence of a test ligand which binds the target protein, than in the absence of a ligand. Binding of the ligand to the target protein can be determined by any method which distinguishes between the folded and unfolded states of the target protein. The function of the target protein need not be known in order for this assay to be performed. Virtually any agent can be assessed by this method as a test ligand, including, but not limited to, metals, polypeptides, proteins, lipids, polysaccharides, polynucleotides and small organic molecules.

Another method for identifying ligands of a target protein is described in Wieboldt *et al.*, *Anal. Chem.*, 69:1683-1691 (1997), incorporated herein by reference. This technique screens combinatorial libraries of 20-30 agents at a time in solution phase for binding to the target protein. Agents that bind to the target

protein are separated from other library components by simple membrane washing. The specifically selected molecules that are retained on the filter are subsequently liberated from the target protein and analyzed by HPLC and pneumatically assisted electrospray (ion spray) ionization mass spectroscopy. This procedure selects library components with the greatest affinity for the target protein, and is particularly useful for small molecule libraries.

In preferred embodiments of the invention, methods of screening for compounds which modulate protease activity comprise contacting test compounds with protease polypeptides and assaying for the presence of a complex between the compound and the protease polypeptide. In such assays, the ligand is typically labelled. After suitable incubation, free ligand is separated from that present in bound form, and the amount of free or uncomplexed label is a measure of the ability of the particular compound to bind to the protease polypeptide.

In another embodiment of the invention, high throughput screening for compounds having suitable binding affinity to protease polypeptides is employed. Briefly, large numbers of different small peptide test compounds are synthesised on a solid substrate. The peptide test compounds are contacted with the protease polypeptide and washed. Bound protease polypeptide is then detected by methods well known in the art. Purified polypeptides of the invention can also be coated directly onto plates for use in the aforementioned drug screening techniques. In addition, non-neutralizing antibodies can be used to capture the protein and immobilize it on the solid support.

Other embodiments of the invention comprise using competitive screening assays in which neutralizing antibodies capable of binding a polypeptide of the invention specifically compete with a test compound for binding to the polypeptide. In this manner, the antibodies can be used to detect the presence of any peptide that shares one or more antigenic determinants with a protease polypeptide. Radiolabeled competitive binding studies are described in A.H. Lin *et al.* *Antimicrobial Agents and Chemotherapy*, 1997, vol. 41, no. 10. pp. 2127-2131, the disclosure of which is incorporated herein by reference in its entirety.

Therapeutic Methods

The invention includes methods for treating a disease or disorder by administering to a patient in need of such treatment a protease polypeptide substantially identical to an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, SEQ ID NO:70, and any other protease polypeptide of the present invention. As discussed in the section "Gene Therapy," a protease polypeptide of the invention may also be administered indirectly by via administration of suitable polynucleotide means for *in vivo* expression of the protease polypeptide. Preferably the protease polypeptide will have 90%, 91%, 92%, 93%, 94%, 95%, 96%, 97%, 98%or 99% identity to one of the aforementioned sequences.

In another aspect, the invention provides methods for treating a disease or disorder by administering to a patient in need of such treatment a substance that modulates the activity of a protease substantially identical to a sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

Preferably the disease is selected from the group consisting of cancers, immune-related diseases and disorders, cardiovascular disease, brain or neuronal-

associated diseases, and metabolic disorders. More specifically these diseases include cancer of tissues, blood, or hematopoietic origin, particularly those involving breast, colon, lung, prostate, cervical, brain, ovarian, bladder, or kidney; central or peripheral nervous system diseases and conditions including migraine, pain, sexual dysfunction, mood disorders, attention disorders, cognition disorders, hypotension, and hypertension; psychotic and neurological disorders, including anxiety, schizophrenia, manic depression, delirium, dementia, severe mental retardation and dyskinesias, such as Huntington's disease or Tourette's Syndrome; neurodegenerative diseases including Alzheimer's, Parkinson's, Multiple sclerosis, and Amyotrophic lateral sclerosis; viral or non-viral infections caused by HIV-1, HIV-2 or other viral- or prion-agents or fungal- or bacterial- organisms; metabolic disorders including Diabetes and obesity and their related syndromes, among others; cardiovascular disorders including reperfusion restenosis, coronary thrombosis, clotting disorders, unregulated cell growth disorders, atherosclerosis; ocular disease including glaucoma, retinopathy, and macular degeneration; inflammatory disorders including rheumatoid arthritis, chronic inflammatory bowel disease, chronic inflammatory pelvic disease, multiple sclerosis, asthma, osteoarthritis, psoriasis, atherosclerosis, rhinitis, autoimmunity, and organ transplant rejection.

In preferred embodiments, the invention provides methods for treating or preventing a disease or disorder by administering to a patient in need of such treatment a substance that modulates the activity of a protease polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. Preferably the disease is selected from the group consisting of cancers, immune-

related diseases and disorders, cardiovascular disease, brain or neuronal-associated diseases, and metabolic disorders. More specifically these diseases include cancer of tissues, blood, or hematopoietic origin, particularly those involving breast, colon, lung, prostate, cervical, brain, ovarian, bladder, or kidney; central or peripheral nervous system diseases and conditions including migraine, pain, sexual dysfunction, mood disorders, attention disorders, cognition disorders, hypotension, and hypertension; psychotic and neurological disorders, including anxiety, schizophrenia, manic depression, delirium, dementia, severe mental retardation and dyskinesias, such as Huntington's disease or Tourette's Syndrome; neurodegenerative diseases including Alzheimer's, Parkinson's, Multiple sclerosis, and Amyotrophic lateral sclerosis; viral or non-viral infections caused by HIV-1, HIV-2 or other viral- or prion-agents or fungal- or bacterial- organisms; metabolic disorders including Diabetes and obesity and their related syndromes, among others; cardiovascular disorders including reperfusion restenosis, coronary thrombosis, clotting disorders, unregulated cell growth disorders, atherosclerosis; ocular disease including glaucoma, retinopathy, and macular degeneration; inflammatory disorders including rheumatoid arthritis, chronic inflammatory bowel disease, chronic inflammatory pelvic disease, multiple sclerosis, asthma, osteoarthritis, psoriasis, atherosclerosis, rhinitis, autoimmunity, and organ transplant rejection.

The invention also features methods of treating or preventing a disease or disorder by administering to a patient in need of such treatment a substance that modulates the activity of a protease polypeptide having an amino acid sequence selected from the group consisting those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70. Preferably the disease

is selected from the group consisting of immune-related diseases and disorders, cardiovascular disease, and cancer. Most preferably, the immune-related diseases and disorders are selected from the group consisting of rheumatoid arthritis, chronic inflammatory bowel disease, chronic inflammatory pelvic disease, multiple sclerosis, asthma, osteoarthritis, psoriasis, atherosclerosis, rhinitis, autoimmunity, and organ transplantation.

Substances useful for treatment of protease-related disorders or diseases preferably show positive results in one or more *in vitro* assays for an activity corresponding to treatment of the disease or disorder in question (Examples of such assays are provided herein, including Example 7). Examples of substances that can be screened for favorable activity are provided and referenced throughout the specification, including this section (**Screening Methods to Identify Substances that Modulate Protease Activity**). The substances that modulate the activity of the proteases preferably include, but are not limited to, antisense oligonucleotides, ribozymes, and other inhibitors of proteases, as determined by methods and screens referenced this section and in Example 7, below, and any other suitable methods. The use of antisense oligonucleotides and ribozymes are discussed more fully in the Section "Gene Therapy," below.

The term "preventing" refers to decreasing the probability that an organism contracts or develops an abnormal condition.

The term "treating" refers to having a therapeutic effect and at least partially alleviating or abrogating an abnormal condition in the organism.

The term "therapeutic effect" refers to the inhibition or activation factors causing or contributing to the abnormal condition. A therapeutic effect relieves to some extent one or more of the symptoms of the abnormal condition. In reference to the treatment of abnormal conditions, a therapeutic effect can refer to one or more of the following: (a) an increase or decrease in the proliferation, growth, and/or differentiation of cells; (b) activation or inhibition (*i.e.*, slowing or stopping) of cell death; (c) inhibition of degeneration; (d) relieving to some extent one or more of the symptoms associated with the abnormal condition; and (e) enhancing the function of

the affected population of cells. Compounds demonstrating efficacy against abnormal conditions can be identified as described herein.

The term "abnormal condition" refers to a function in the cells or tissues of an organism that deviates from their normal functions in that organism. An abnormal condition can relate to cell proliferation, cell differentiation, or cell survival.

Abnormal cell proliferative conditions include cancers such as fibrotic and mesangial disorders, abnormal angiogenesis and vasculogenesis, wound healing, psoriasis, diabetes mellitus, and inflammation.

Abnormal differentiation conditions include, but are not limited to neurodegenerative disorders, slow wound healing rates, and slow tissue grafting healing rates.

Abnormal cell survival conditions relate to conditions in which programmed cell death (apoptosis) pathways are activated or abrogated. A number of proteases are associated with the apoptosis pathways. Aberrations in the function of any one of the proteases could lead to cell immortality or premature cell death.

The term "aberration", in conjunction with the function of a protease in a signal transduction process, refers to a protease that is over- or under-expressed in an organism, mutated such that its catalytic activity is lower or higher than wild-type protease activity, mutated such that it can no longer interact with a natural binding partner, is no longer modified by another protein, or no longer interacts with a natural binding partner.

The term "administering" relates to a method of incorporating a compound into cells or tissues of an organism. The abnormal condition can be prevented or treated when the cells or tissues of the organism exist within the organism or outside of the organism. Cells existing outside the organism can be maintained or grown in cell culture dishes. For cells harbored within the organism, many techniques exist in the art to administer compounds, including (but not limited to) oral, parenteral, dermal, injection, and aerosol applications. For cells outside of the organism, multiple techniques exist in the art to administer the compounds, including (but not

limited to) cell microinjection techniques, transformation techniques, and carrier techniques.

The abnormal condition can also be prevented or treated by administering a compound to a group of cells having an aberration in a signal transduction pathway to an organism. The effect of administering a compound on organism function can then be monitored. The organism is preferably a mouse, rat, rabbit, guinea pig, or goat, more preferably a monkey or ape, and most preferably a human.

In another aspect, the invention features methods for detection of a protease polypeptide in a sample as a diagnostic tool for diseases or disorders, wherein the method comprises the steps of: (a) contacting the sample with a nucleic acid probe which hybridizes under hybridization assay conditions to a nucleic acid target region of a protease polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, said probe comprising the nucleic acid sequence encoding the polypeptide, fragments thereof, and the complements of the sequences and fragments; and (b) detecting the presence or amount of the probe:target region hybrid as an indication of the disease.

In preferred embodiments of the invention, the disease or disorder is selected from the group consisting of rheumatoid arthritis, arteriosclerosis, autoimmune disorders, organ transplantation, myocardial infarction, cardiomyopathies, stroke, renal failure, oxidative stress-related neurodegenerative disorders, and cancer. Preferably the disease is selected from the group consisting of cancers, immune-related diseases and disorders, cardiovascular disease, brain or neuronal-associated diseases, and metabolic disorders. More specifically these diseases include cancer

of tissues, blood, or hematopoietic origin, particularly those involving breast, colon, lung, prostate, cervical, brain, ovarian, bladder, or kidney; central or peripheral nervous system diseases and conditions including migraine, pain, sexual dysfunction, mood disorders, attention disorders, cognition disorders, hypotension, and hypertension; psychotic and neurological disorders, including anxiety, schizophrenia, manic depression, delirium, dementia, severe mental retardation and dyskinesias, such as Huntington's disease or Tourette's Syndrome; neurodegenerative diseases including Alzheimer's, Parkinson's, Multiple sclerosis, and Amyotrophic lateral sclerosis; viral or non-viral infections caused by HIV-1, HIV-2 or other viral- or prion-agents or fungal- or bacterial- organisms; metabolic disorders including Diabetes and obesity and their related syndromes, among others; cardiovascular disorders including reperfusion restenosis, coronary thrombosis, clotting disorders, unregulated cell growth disorders, atherosclerosis; ocular disease including glaucoma, retinopathy, and macular degeneration; inflammatory disorders including rheumatoid arthritis, chronic inflammatory bowel disease, chronic inflammatory pelvic disease, multiple sclerosis, asthma, osteoarthritis, psoriasis, atherosclerosis, rhinitis, autoimmunity, and organ transplant rejection.

The protease "target region" is the nucleotide base sequence selected from the group consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35, or the corresponding full-length sequences, a functional derivative thereof, or a fragment thereof or a domain thereof to which the nucleic acid probe will specifically hybridize. Specific hybridization indicates that in the presence of other nucleic acids the probe only hybridizes detectably with the nucleic acid target region of the protease of the invention. Putative target regions

can be identified by methods well known in the art consisting of alignment and comparison of the most closely related sequences in the database.

In preferred embodiments the nucleic acid probe hybridizes to a protease target region encoding at least 6, 12, 75, 90, 105, 120, 150, 200, 250, 300 or 350 contiguous amino acids of a sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or the corresponding full-length amino acid sequence, or a functional derivative thereof. Hybridization conditions should be such that hybridization occurs only with the protease genes in the presence of other nucleic acid molecules. Under stringent hybridization conditions only highly complementary nucleic acid sequences hybridize. Preferably, such conditions prevent hybridization of nucleic acids having more than 1 or 2 mismatches out of 20 contiguous nucleotides. Such conditions are defined in Berger *et al.* (1987) (Guide to Molecular Cloning Techniques pg 421, hereby incorporated by reference herein in its entirety including any figures, tables, or drawings.).

The diseases for which detection of protease genes in a sample could be diagnostic include diseases in which protease nucleic acid (DNA and/or RNA) is amplified in comparison to normal cells. By "amplification" is meant increased numbers of protease DNA or RNA in a cell compared with normal cells. In normal cells, proteases may be found as single copy genes. In selected diseases, the chromosomal location of the protease genes may be amplified, resulting in multiple copies of the gene, or amplification. Gene amplification can lead to amplification of protease RNA, or protease RNA can be amplified in the absence of protease DNA amplification.

"Amplification" as it refers to RNA can be the detectable presence of protease RNA in cells, since in some normal cells there is no basal expression of protease RNA. In other normal cells, a basal level of expression of protease exists, therefore in these cases amplification is the detection of at least 1-2-fold, and preferably more, protease RNA, compared to the basal level.

The diseases that could be diagnosed by detection of protease nucleic acid in a sample preferably include cancers. The test samples suitable for nucleic acid probing methods of the present invention include, for example, cells or nucleic acid extracts of cells, or biological fluids. The samples used in the above-described methods will vary based on the assay format, the detection method and the nature of the tissues, cells or extracts to be assayed. Methods for preparing nucleic acid extracts of cells are well known in the art and can be readily adapted in order to obtain a sample that is compatible with the method utilized.

In a final aspect, the invention features a method for detection of a protease polypeptide in a sample as a diagnostic tool for a disease or disorder, wherein the method comprises: (a) comparing a nucleic acid target region encoding the protease polypeptide in a sample, where the protease polypeptide has an amino acid sequence selected from the group consisting those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or one or more fragments thereof, with a control nucleic acid target region encoding the protease polypeptide, or one or more fragments thereof; and (b) detecting differences in sequence or amount between the target region and the control target region, as an indication of the disease or disorder. Preferably the disease is selected from the

group consisting of cancers, immune-related diseases and disorders, cardiovascular disease, brain or neuronal-associated diseases, and metabolic disorders.

More specifically these diseases include cancer of tissues, blood, or hematopoietic origin, particularly those involving breast, colon, lung, prostate, cervical, brain, ovarian, bladder, or kidney; central or peripheral nervous system diseases and conditions including migraine, pain, sexual dysfunction, mood disorders, attention disorders, cognition disorders, hypotension, and hypertension; psychotic and neurological disorders, including anxiety, schizophrenia, manic depression, delirium, dementia, severe mental retardation and dyskinesias, such as Huntington's disease or Tourette's Syndrome; neurodegenerative diseases including Alzheimer's, Parkinson's, Multiple sclerosis, and Amyotrophic lateral sclerosis; viral or non-viral infections caused by HIV-1, HIV-2 or other viral- or prion-agents or fungal- or bacterial- organisms; metabolic disorders including Diabetes and obesity and their related syndromes, among others; cardiovascular disorders including reperfusion restenosis, coronary thrombosis, clotting disorders, unregulated cell growth disorders, atherosclerosis; ocular disease including glaucoma, retinopathy, and macular degeneration; inflammatory disorders including rheumatoid arthritis, chronic inflammatory bowel disease, chronic inflammatory pelvic disease, multiple sclerosis, asthma, osteoarthritis, psoriasis, atherosclerosis, rhinitis, autoimmunity, and organ transplant rejection.

The term "comparing" as used herein refers to identifying discrepancies between the nucleic acid target region isolated from a sample, and the control nucleic acid target region. The discrepancies can be in the nucleotide sequences, *e.g.* insertions, deletions, or point mutations, or in the amount of a given nucleotide sequence. Methods to determine these discrepancies in sequences are well-known to one of ordinary skill in the art. The "control" nucleic acid target region refers to the sequence or amount of the sequence found in normal cells, *e.g.* cells that are not diseased as discussed previously.

The term "domain" refers to a region of a polypeptide which serves a particular function. For instance, N-terminal or C-terminal domains of signal

transduction proteins can serve functions including, but not limited to, binding molecules that localize the signal transduction molecule to different regions of the cell or binding other signaling molecules directly responsible for propagating a particular cellular signal. Some domains can be expressed separately from the rest of the protein and function by themselves, while others must remain part of the intact protein to retain function. The latter are termed functional regions of proteins and also relate to domains.

The expression of proteases can be modulated by signal transduction pathways such as the Ras/MAP kinase signaling pathways. Additionally, the activity of proteases can modulate the activity of the MAP kinase signal transduction pathway. Furthermore, proteases can be shown to be instrumental in the communication between disparate signal transduction pathways.

The term "signal transduction pathway" refers to the molecules that propagate an extracellular signal through the cell membrane to become an intracellular signal. This signal can then stimulate a cellular response. The polypeptide molecules involved in signal transduction processes are typically receptor and non-receptor protein tyrosine kinases, receptor and non-receptor protein phosphatases, polypeptides containing SRC homology 2 and 3 domains, phosphotyrosine binding proteins (SRC homology 2 (SH2) and phosphotyrosine binding (PTB and PH) domain containing proteins), proline-rich binding proteins (SH3 domain containing proteins), GTPases, phosphodiesterases, phospholipases, prolyl isomerases, proteases, Ca^{2+} binding proteins, cAMP binding proteins, guanyl cyclases, adenylyl cyclases, NO generating proteins, nucleotide exchange factors, and transcription factors.

The summary of the invention described above is not limiting and other features and advantages of the invention will be apparent from the following detailed description of the invention, and from the claims.

BRIEF DESCRIPTION OF THE FIGURES

Figures 1A-W shows the partial nucleotide sequences for human proteases oriented in a 5' to 3' direction (SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35). In the sequences, N means any nucleotide.

Figure 2A-I shows the partial amino acid sequences for the human proteases encoded by SEQ ID No. 1-35 in the direction of translation (SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70). In the sequences, X means any amino acid.

DETAILED DESCRIPTION OF THE INVENTION

The following description of the background of the invention is provided to aid in understanding the invention, but is not admitted to be or to describe prior art to the invention.

Proteases are enzymes capable of severing the amino acid backbone of other proteins, and are involved in a large number of diverse processes within the body. Their normal functions include modulation of apoptosis (caspases) (Salvesen and Dixon, *Cell*, 1997, 91:443-46), control of blood pressure (renin, angiotensin-

converting enzymes) (van Hooft *et al.*, 1991, N Engl J Med. 324(19):1305-11, and chapters 254 and 359 in Barrett *et al.*, Handbook of Proteolytic Enzymes, 1998, Academic Press, San Diego), tissue remodeling and tumor invasion (collagenase) (Vu *et al.*, 1998, *Cell* 93:411-22, Werb, 1997, *Cell*, 91:439-442), development of Alzheimer's Disease (β-secretase) (De Strooper *et al.*, 1999, *Nature* 398:518-22), protein turnover and cell-cycle regulation (proteosome) (Bastians *et al.*, 1999, *Mol. Biol. Cell.* 10:3927-41, Gottesman, *et al.*, 1997, *Cell*, 91:435-38, Larsen *et al.*, 1997, *Cell*, 91:431-34), inflammation (TNF-α convertase) (Black *et al.*, *Nature*, 1997, 385:729-33), and protein turnover (Bochtler *et al.*, 1999, *Annu. Rev. Biophys Biomol Struct.*28:295-317). Proteases may be classified into several major groups including serine proteases, cysteine proteases, aspartyl proteases, metalloproteases, threonine proteases, and other proteases.

1. Aspartyl proteases (A1; Prosite number PS00141):

Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. Examples of aspartyl protease polypeptides according to the invention include SGPr140, r197, r005 and r078 (SEQ ID NOS:1, 2, 3, and 4, respectively). These polypeptides may have one or more of the following activites.

Cathepsins

Cathepsin E is an immunologically discrete aspartic protease found in the gastrointestinal tract (Azuma *et al.*, 1992, *J. Biol. Chem.*, 267:1609-1614). Cathepsin E is an intracellular proteinase that does not appear to be involved in the digestion of dietary protein. It is found in highest concentration in the surface of epithelial mucus-producing cells of the stomach. It is the first aspartic proteinase expressed in the fetal stomach and is found in more than half of gastric cancers. It appears, therefore, to be an 'oncofetal' antigen. Its association with stomach cancers suggests it may play a role in the development of this disease.

Cathepsin D, a lyosomal aspartyl protease, is being studied as a prognostic marker in various cancers, in particular, breast cancer. (Rochefort *et al.*, *Clin. Chim. Acta*, 2000, 291:157-170).

Renin

Released by the juxtaglomerular cells of the kidney, renin catalyzes the first step in the activation pathway of angiotensinogen--a cascade that can result in aldosterone release, vasoconstriction, and increase in blood pressure. Renin cleaves angiotensinogen to form angiotensin I, which is converted to angiotensin II by angiotensin I converting enzyme, an important regulator of blood pressure and electrolyte balance. Renin occurs in other organs than the kidney, *e.g.*, in the brain, where it is implicated in the regulation of numerous activities.

Presenilin proteins

Alzheimer's disease (AD) patients with an inherited form of the disease carry mutations in the presenilin proteins (PSEN1; PSEN2) or the amyloid precursor protein (APP). These disease-linked mutations result in increased production of the longer form of amyloid-beta (main component of amyloid deposits found in AD brains) (Saftig *et al.*, *Eur. Arch. Psychiartry Clin. Neurosci.*, 1999, 249:271-79). Presenilins are postulated to regulate APP processing through their effects on γ-secretase, an enzyme that cleaves APP (Cruts *et al.*, 1998, *Hum. Mutat.*, 11:183-190, Haass *et al.*, *Science*, 1999, 286:916-19). Also, it is thought that the presenilins are involved in the cleavage of the Notch receptor, such that that they either directly regulate γ-secretase activity or themselves are protease enzymes (De Strooper *et al.*, *Nature*, 1999, 398:518-22). Two alternative transcripts of PSEN2 have been identified (Sato *et al.*, 1999, *J Neurochem.* 72(6):2498-505). Point mutations in the PS1 gene result in a selective increase in the production of the amyloidogenic peptide amyloid-beta (1-42) by proteolytic processing of the amyloid precursor protein (APP)(Lemere *et al.*, 1996, *Nat. Med.* 2(10):1146-50). The possible role of PS1 in normal APP processing was studied by De Strooper *et al.* (*Nature* 391: 387-

390, 1998) in neuronal cultures derived from PS1-deficient mouse embryos. They found that cleavage by α- and β-secretase of the extracellular domain of APP was not affected by the absence of PS1, whereas cleavage by γ-secretase of the transmembrane domain of APP was prevented, causing C-terminal fragments of APP to accumulate and a 5-fold drop in the production of amyloid peptide. Pulse-chase experiments indicated that PS1 deficiency specifically decreased the turnover of the membrane-associated fragments of APP. Thus, PS1 appears to facilitate a proteolytic activity that cleaves the integral membrane domain of APP. The results indicated to the authors that mutations in PS1 that manifest clinically cause a gain of function, and that inhibition of PS1 activity is a potential target for anti-amyloidogenic therapy in Alzheimer disease.

β-secretase

β-secretase, expressed specifically in the brain, is responsible for the proteolytic processing of the amyloid precursor protein (APP) associated with Alzheimer's disease (Potter *et al.*, 2000, *Nat. Biotechnol.* 18(2):125-26). It cleaves at the amino terminus of the β-peptide sequence, between residues 671 and 672 of APP, leading to the generation and extracellular release of β-cleaved soluble APP, and a carboxyterminal fragment that is later released by γ-secretase (Kimberly *et al.*, 2000, *J. Biol. Chem.* 275(5):3173-78). Yan *et al.* (*Nature*, 1999 402:533-37) identified a new membrane-bound aspartyl protease (Asp2) with β-secretase activity. The Asp2 gene is expressed widely in brain and other tissues. Decreasing the expression of Asp2 in cells reduces amyloid β-peptide production and blocks the accumulation of the carboxy-terminal APP fragment that is created by β-secretase cleavage. Asp2 is a new protein target for drugs that are designed to block the production of amyloid β-peptide peptide and the consequent formation of amyloid plaque in Alzheimer's disease.

Two aspartyl proteases involved in human placentation have recently been isolated: decidual aspartyl protease (DAP-1), and DAP-2. (Moses *et al., Mol. Hum. Reprod.*, 1999, 5:983-89)

Another member of the aspartyl peptidase family is HIV-1 retropepsin, from the human immunodeficiency virus type 1. This enzyme is vital for processing of the viral polyprotein and maturation of the mature virion.

2. Cysteine proteases

Another class of proteases which perform a wide variety of functions within the body are the cysteine proteases. Among their roles are the processing of precursor proteins, and intracellular degradation of proteins marked for disposal via the ubiquitin pathway. Catalysis proceeds through a thioester intermediate and is facilitated by a nearby histidine side chain; an asparagine completes the essential catalytic triad. Peptidases in this family with important roles in disease include calpain, the caspases, hedgehog, papain, and Ubiquitin hydrolases. Examples of cysteine protease polypeptides of the present invention include SGPr084, r009, r286, r008, r198, r210, r290, r116, r003, r016 (SEQ ID NOS:5, 6, 7, 8, 9, 10, 11, 12, and 13, respectively). These polypeptides may have one or more of the following activities.

Cysteine proteases are produced by a large number of cells including those of the immune system (macrophages, monocytes, etc.). These immune cells exercise their protective role in the body, in part, by migrating to sites of inflammation and secreting molecules, among the secreted molecules are cysteine proteases.

Under some conditions, the inappropriate regulation of cysteine proteases of the immune system can lead to autoimmune diseases such as rheumatoid arthritis. For example, the over-secretion of the cysteine protease cathepsin C causes the degradation of elastin, collagen, laminin and other structural proteins found in bones. Bone subjected to this inappropriate digestion is more susceptible to metastasis.

Cysteine proteases may also influence vascular permeability through their effect on the kallikrein/kinin pathway, their ability to form complexes with hemagglutinins, their effect in activation of complement components and their ability to destroy serpins.

Caspase (C14) - apopotosis

A cascade of protease reactions is believed to be responsible for the apoptotic changes observed in mammalian cells undergoing programmed cell death. This cascade involves many members of the aspartate-specific cysteine proteases of the caspase family, including Caspases 2, 3, 6, 7, 8, and 10 ((Salvesen and Dixit, *Cell*, 1997, 91:443-446). Cancer cells that escape apoptotic signals, generated by cytotoxic chemotherapeutics or loss of normal cellular survival signals (as in metastatic cells), can go on to develop palpable tumors.

Other caspases are also involved in the activation of pro-inflammatory cytokines. Caspase 1 specifically processes the precursors of IL-1β, and IL-18 (interferon-γ-inducing factor) (Salvesen and Dixit, *Cell*, 1997).

Calpain (C2) - axonal death, dystrophies

Calcium-dependent cysteine proteases, collectively called calpain, are widely distributed in mammalian cells (Wang, 2000, *Trends Neurosci.* 23(1):20-26). The calpains are nonlysosomal intracellular cysteine proteases. The mammalian calpains include 2 ubiquitous proteins, CAPN1 and CAPN2, as well as 2 stomach-specific proteins, and CAPN3, which is muscle-specific (Herasse *et al.*, 1999, *Mol. Cell. Biol.* 19(6):4047-55). The ubiquitous enzymes consist of heterodimers with distinct large subunits associated with a common small subunit, all of which are encoded by different genes. The large subunits of calpains can be subdivided into 4 domains; domains I and III, whose functions remain unknown, show no homology with known proteins. The former, however, may be important for the regulation of the proteolytic activity. Domain II shows similarity with other cysteine proteases, which share histidine, cysteine, and asparagine residues at their active sites. Domain

IV is calmodulin-like. CAPN5 and CAPN6 differ from previously identified vertebrate calpains in that they lack a calmodulin-like domain IV (Ohno *et al.*, 1990, *Cytogenet. Cell Genet.* 53(4):225-29).

Mutations in the CAPN3 gene have been associated with limb-girdle muscular dystrophy, type 2A (LGMD2A) (Allamand *et al.*, 1995, *Hum. Molec. Genet.* 4:459-463). The slowly progressive muscle weakness associated with this disease is usually first evident in the pelvic girdle and then spreads to the upper limbs while sparing facial muscles. Calpain has also been implicated in the development of hyperactive Cdk5 leading to neuronal cell death associated with Alzheimer's disease (Patrick *et al.*, 1999, *Nature* 402:615-622).

Hedgehog (C46) – Cancer

The organization and morphology of the developing embryo are established through a series of inductive interactions. One family of vertebrate genes has been described related to the Drosophila gene 'hedgehog' (hh) that encodes inductive signals during embryogenesis (Johnson and Tabin, 1997, *Cell* 90:979-990). 'Hedgehog' encodes a secreted protein that is involved in establishing cell fates at several points during Drosophila development (Marigo *et al.*, 1995, *Genomics* 28:44-51). There are 3 known mammalian homologs of hh: Sonic hedgehog (Shh), Indian hedgehog (Ihh), and desert hedgehog (Dhh) (Johnson and Tabin, 1997, *Cell* 90:979-990). Like its Drosophila cognate, Shh encodes a signal that is instrumental in patterning the early embryo. It is expressed in Hensen's node, the floorplate of the neural tube, the early gut endoderm, the posterior of the limb buds, and throughout the notochord (Chiang *et al.*, 1996, *Nature* 383:407-413). It has been implicated as the key inductive signal in patterning of the ventral neural tube, the anterior-posterior limb axis, and the ventral somites. Oro *et al.* ("Basal cell carcinomas in mice overexpressing sonic hedgehog." *Science* 276: 817-821, 1997) showed that transgenic mice overexpressing SHH in the skin developed many features of the basal cell nevus syndrome, demonstrating that SHH is sufficient to induce basal cell carcinomas (BCCs) in mice. The data suggested that SHH may

have a role in human tumorigenesis. Activating mutations of SHH or another 'hedgehog' gene may be an alternative pathway for BCC formation in humans. The human mutation his133tyr (his134tyr in mouse) is a candidate. It is distinct from loss-of-function mutations reported for individuals with holoprosencephaly (Oro *et al.*, 1997, *Science* 276:817-821). His133 lies adjacent in the catalytic site to his134, one of the conserved residues thought to be necessary for catalysis. SHH may be a dominant oncogene in multiple human tumors, a mirror of the tumor suppressor activity of the opposing 'patched' (PTCH) gene (Aszterbaum *et al.*, 1998, *J. Invest. Derm.* 110:885-888). The rapid and frequent appearance of Shh-induced tumors in the mice suggested that disruption of the SHH-PTC pathway is sufficient to create BCCs.

Members of the vertebrate hedgehog family (Sonic, Indian, and Desert) have been shown to be essential for the development of various organ systems, including neural, somite, limb, skeletal, and for male gonad morphogenesis. Desert hedgehog is expressed in the developing retina, whereas Indian hedgehog (Ihh) is expressed in the developing and mature retinal pigmented epithelium beginning at embryonic day 13 (Levine *et al.*, *J.Neurosci.*, 1997, 17(16):6277-88). Dhh has also been implicated in having a role in the regulation of spermatogenesis. Sertoli cell precursors express Sry, sex determining gene, which leads to testis development in mammals. Dhh expression is initiated in Sertoli cell precursors shortly after the activation of Sry and persists in the testis into the adult. Bitgood *et al.* (*Curr. Biol.*, 1996, 6(3):298-304) disclose that female mice homozygous for a Dhh-null mutation show no obvious phenotype, whereas males are viable but infertile having a complete absence of mature sperm, demonstrating that Dhh signaling plays an essential role in the regulation of mammalian spermatogenesis. Dhh has also been found to have a role in the and maintenance of protective nerve sheaths endo-, peri- and epineurium. In Dhh knockout mice, the connective tissue sheaths in adult nerves appear highly abnormal by electron microscopy. Mirsky *et al.*, (*Ann. N.Y. Acad. Sci.*, 1999, 883:196-202) demonstrate that Dhh signaling from Schwann cells to the

mesenchyme is involved in the formation of a morphologically and functionally normal perineurium.

Recent advances in developmental and molecular biology during embryogenesis and organogenesis have provided new insights into the mechanism of bone formation. Iwasaki *et al.*, (*J. Bone Joint Surg. Br.*, 1999, 81(6):1076-82) demonstrate that Indian Hedgehog (Ihh) is expressed in cartilage cell precursors and later in mature and hypertrophic chondrocytes. Ihh plays a critical role in the morphogenesis of the vertebrate skeleton. Becker *et al.* (*Dev. Biol.*, 1997, 187(2):298-310) provide data which suggests that Ihh is also involved in mediating differentiation of extraembryonic endoderm during early mouse embryogenesis. Short limbed dwarfism, with decreased chondrocyte proliferation and extensive hypertrophy are the results of targeted deletion of Ihh (Karp *et al.*, 2000, Development 127(3):543-48). The expression of Ihh mRNA and protein is unregulated dramatically as F9 cells differentiate in response to retinoic acid, into either parietal endoderm or embryoid bodies, containing an outer visceral endoderm layer. RT-PCR analysis of blastocyst outgrowth cultures demonstrates that whereas little or no Ihh message is present in blastocysts, significant levels appear upon subsequent days of culture, coincident with the emergence of parietal endoderm cells.

Ubiquitin hydrolases (C12) - apoptosis, checkpoint integrity

Ubiquitin carboxyl-terminal hydrolases (3.1.2.15) (deubiquitinating enzymes) are thiol proteases that recognize and hydrolyze the peptide bond at the C-terminal glycine of ubiquitin. These enzymes are involved in the processing of poly-ubiquitin precursors as well as that of ubiquinated proteins. In eukaryotic cells, the covalent attachment of ubiquitin to proteins plays a role in a variety of cellular processes. In many cases, ubiquitination leads to protein degradation by the 26S proteasome. Protein ubiquitination is reversible, and the removal of ubiquitin is catalyzed by deubiquitinating enzymes, or DUBs. A defect in these enzymes, catalyzing the removal of ubiquitin from ubiquinated proteins, may be characteristic

of neurodegenerative diseases such as Alzheimer's, Parkinson's, progressive supranuclear palsy, and Pick's and Kuf's disease.

Papain (C1) – cathepsins K, S and B,- bone resorbtion, Ag processing (Prosite PS00139)

Cathepsin K, a member of the papain family of peptidases, is involved in osteoclastic resorption. It plays an important role in extracellular degradation and may have a role in disorders of bone remodeling, such as pyncodysostosis, an autosomal recessive osteochondrodysplasia characterized by osteosclerosis and short stature. Antigen presentation by major histocompatibility complex (MHC) class II molecules requires the participation of different proteases in the endocytic route to degrade endocytosed antigens as well as the MHC class II-associated invariant chain. Only cathepsin S, a member of the papain family, appears to be essential for complete destruction of the invariant chain. Cathepsin B is overexpressed in tumors of the lung, prostate, colon, breast, and stomach. Hughes *et al. (Proc. Nat. Acad. Sci.* 95: 12410-12415, 1998) found an amplicon at 8p23-p22 that resulted in cathepsin B overexpression in esophageal adenocarcinoma. Abundant extracellular expression of CTSB protein was found in 29 of 40 (72.5%) of esophageal adenocarcinoma specimens by use of immunohistochemical analysis. The findings were thought to support an important role for CTSB in esophageal adenocarcinoma and possibly in other tumors.

Cathepsin B, a lyosomal protease, is being studied as a prognostic marker in various cancers (breast, pulmonary adenocarcinomas).

Cysteine Protease AEP

The cysteine protease AEP plays another role in the immune functions. It has been implicated in the protease step required for antigen processing in B cells. (Manoury *et al. Nature* 396:695-699 (1998))

Hepatitis A viral protease (C3E)

The Hepatitis A genome encodes a cysteine protease required for enzymatic cleavages *in vivo* to yield mature proteins (Wang, 1999, *Prog. Drug Res.* 52:197-219). This enzyme and its homologs in other viruses (such as hepatitis E virus) are potential targets for chemotherapeutic intervention.

3. Metalloproteases

Examples of metalloprotease protease polypeptides according to the invention include SGPr016, r352, r050, r282, r046, r060, r068, r096, r119, r143, r164, r281, r075, r292, r069, r212, r049, r026, r203, r157, r154, r088 (SEQ ID NOS:14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, and 35 respectively). These polypeptides may have one or more of the following activities.

Collagenase (M10) - invasion

Matrix degradation is an essential step in the spread of cancer. The 72- and 92-kD type IV collagenases are members of a group of secreted zinc metalloproteases which, in mammals, degrade the collagens of the extracellular matrix. Other members of this group include interstitial collagenase and stromelysin (Nagase *et al.*, 1992, *Matrix Suppl.* 1:421-424). By targeted disruption in embryonic stem cells, Vu *et al.* (*Cell*, 1998, 934:11-22) created homozygous mice with a null mutation in the MMP9/gelatinase B gene. These mice exhibited an abnormal pattern of skeletal growth plate vascularization and ossification. Growth plates from MMP9-null mice in culture showed a delayed release of an angiogenic activator, establishing a role for this proteinase in controlling angiogenesis.

MMP2 (gelatinase A) have been associated with the aggressiveness of human cancers (Chenard *et al.*, 1999, *Int. J. Cancer*, 82:208-12). In a study comparing basal cell carcinomas (BCC) with the more aggressive squamous cell carcinomas (SCC), both MMP2 and MMP9 were expressed at a higher level in SCC (Dumas *et al.*, 1999, *Anticancer Res.*, 19(4B):2929-38). Additionally, expression of

MMP2 and MMP9 in T lymphocytes has recently been shown to be modulated by the Ras/MAP kinase signaling pathways (Esparza *et al.*, 1999, *Blood*, 94:2754-66) (see also, Li *et al.*, 1998, *Biochim. Biophys. Acta*, 1405:110-20).

ADAMs (M12) - TNF, inflammation, growth factor processing

The ADAM peptidases are a family of proteins containing a disintegrin and metalloproteinase (ADAM) domain (Werb and Yan, *Science*, 1998, 282:1279-1280). Members of this family are cell surface proteins with a unique structure possessing both potential adhesion and protease domains (Primakoff and Myles, *Trends in Genet.*, 2000, 16:83-87). Activity of these proteases can be linked to TNF, inflammation, and/or growth factor processing.

ADAM proteases have also been characterized as having a pro- and metalloproteinase domain, a disintegrin domain, a cysteine-rich region and an EGF repeat (Blobel, 1997, *Cell*, 90:589-592 which is hereby incorporated herein by reference in its entirety including any figures, tables, or drawings). They have been associated with the release from the plasma membrane of numerous proteins including Tumor Necrosis Factor-α (TNF-α), kit-ligand, TGFα, Fas-ligand, cytokine receptors such as the Il-6 receptor and the NGF receptor, as well as adhesion proteins such as L-selectin, and the b amyloid precursor proteins (Blobel, 1997, *Cell*, 90:589-592).

Tumor necrosis factor-α is synthesized as a proinflammatory cytokine from a 233-amino acid precursor. Conversion of the membrane-bound precursor to a secreted mature protein is mediated by a protease termed TNF-α convertase. TNF-α is involved in a variety of diseases. ADAM17, which contains a disintegrin and metalloproteinase domains, is also called 'tumor necrosis factor-α converting enzyme' (TACE) (Black *et al.*, *Nature*, 1997, 385:729-33). The gene encodes an 824-amino acid polypeptide containing the features of the ADAM family: a secretory signal sequence, a disintegrin domain, and a metalloprotease domain. Expression studies showed that the encoded protein cleaves precursor tumor

necrosis factor-α to its mature form. This enzyme may also play a role in the processing of Transforming Growth Factor-α (TGF-α), as mice which lack the gene are similar in phenotype to those that lack TGF-α (Peschon *et al.*, *Science*, 282:1281-1284).

Neprylisin (M13) - Endothelin-converting enzyme

Neprylisin, a metallopeptidase active in degradation of enkephalins and other bioactive peptides, is a drug target in hypertension and renal disease (Oefner, *et al.*, *J. Mol. Biol.*, 2000, 296:341-49).

Carboxypeptidase (M14) - Neurotransmitter processing

Carboxypeptidases specifically remove COOH-terminal basic amino acids(arginine or lysine) (Nesheim, 1998, *Curr. Opin. Hematol.* 5(5):309-13). They have important functions in many biologic processes, including activation, inactivation, or modulation of peptide hormone activity, neurotransmitter processing, and alteration of physical properties of proteins and enzymes (Ostrowska *et al.*, 1998, *Rocz. Akad. Med. Bialymst.* 43:39-55).

Dipeptidase (M2) - ACE

Angiotensin I converting enzyme (EC 3.4.15.1), or kininase II, is adipeptidyl carboxypeptidase that plays an important role in blood pressure regulation and electrolyte balance by hydrolyzing angiotensin I into angiotensin II, a potent vasopressor, andaldosterone-stimulating peptide. The enzyme is also able to inactivate bradykinin, a potent vasodilator. Although angiotensin-converting enzyme has been studied primarily in the context of its role in blood pressure regulation, this widely distributed enzyme has many other physiologic functions. There are two forms of ACE: a testis-specific isozyme and a somatic isozyme which has two active centers.

Matrix metalloproteases (M10B) – tissue remodeling and inflammation

The matrix metalloproteases (MMPs) are a family of related matrix-degrading enzymes that are important in tissue remodeling and repair during development and inflammation (Belotti *et al.*, 1999, *Int. J. Biol. Markers* 14(4):232-38). Abnormal expression is associated with various diseases such as tumor invasiveness (Johansson and Kahari, 2000, *Histol. Histopathol.* 15(1):225-37), arthritis (Malemud *et al.*, 1999, *Front. Biosci.* 4:D762-71), and atherosclerosis (Nagase, 1997, *Biol. Chem.* 378(3-4):151-60). MMP activity may also be related to tobacco-induced pulmonary emphysema (Dhami *et al.*, *Am. J. Respir. Cell Mol. Biol.*, 2000, 22:244-52).

SREBP Protease (M50)

The sterol regulatory element-binding proteins protease functions in the intra-membrane proteolysis and release of sterol-regulatory binding proteins (SREBPs) (Duncan *et al.*, 1997, *J. Biol. Chem.* 272:12778-85). SREBPs activate genes of cholesterol and fatty acid metabolism, making the SREBP protease an attractive target for therapeutic modulation (Brown *et al.*, 1997, *Cell* 89:331-340).

Metalloprotease processing of growth factors

In addition to the processing of TGF-α described above, metalloproteases have been directly demonstrated to be active in the processing of the precursor of other growth factors such as heparin-binding EGF (proHB-EFG) (Izumi *et al.*, *EMBO J*, 1998,17:7260-72), and amphiregulin (Brown *et al.*, 1998, *J. Biol. Chem.*, 27:17258-68).

Additionally, metalloproteases have recently been shown to be instrumental in the communication whereby stimulation of a GPCR pathway results in stimulation of the MAP kinase pathway (Prenzel *et al.*, 1999, *Nature*, 402:884-888). The growth factor intermediate in the pathway, HB-EGF is released by the cell in a proteolytic step regulated by the GPCR pathway involving an uncharacterized metalloprotease. After release, the HB-EGF is bound by the extracellular matrix and

then presented to the EGF receptors on the surface, resulting in the activation of the MAP kinase pathway (Prenzel *et al.*, 1999, *Nature*, 402:884-888).

A recent study by Gallea-Robache *et al.* (1997) has also implicated a metalloprotease family displaying different substrate specificites in the shedding of other growth factors including macrophage colony-stimulating factor (M-CSF) and stem cell factor (SCF) (Gallea-Robache *et al.*, 1997, *Cytokine* 9:340-46). The shedding of M-CSF (also known as CSF-1) has been linked to activation of Protein Kinase C by phorbol esters (Stein *et al.*, 1991, *Oncogene*, 6:601-05).

4. Serine Proteases

The serine proteases are a class which includes trypsin, kallikrein, chymotrypsin, elastase, thrombin, tissue plasminogen activator (tPA), urokinase plasminogen activator (uPA), plasmin (Werb, *Cell*, 1997, 91:439-442), kallikrein (Clements, *Biol. Res.*, 1998, 31151-59), and cathepsin G (Shamamian *et al.*, *Surgery*, 2000, 127:142-47). These proteases have in common a well-conserved catalytic triad of amino acid residues in their active site consisting of histidine-57, aspartic acid-102, and serine-195 (using the chymotrypsin numbering system). Serine protease activity has been linked to coagulation and they may have use as tumor markers.

Serine proteases can be further subclassified by their specificity in substrates. The elastases prefer to cleave substrates adjacent to small aliphatic residues such as valine, chymases prefer to cleave near large aromatic hydrophobic residures, and tryptases prefer positively charged residues. One additional class of serine protease has been described recently which prefers to cleave adjacent to a proline. This prolyl endopeptidase has been implicated in the progression of memory loss in Alzheimer's patients (Toide *et al.*, 1998, *Rev. Neurosci.* 9(1):17-29).

A partial list of proteases known to belong to this large and important family include: blood coagulation factors VII, IX, X, XI and XII; thrombin; plasminogen; complement components C1r, C1s, C2; complement factors B, D and I; complement-activating component of RA-reactive factor; elastases 1, 2, 3A, 3B

(protease E); hepatocyte growth factor activator; glandular (tissue) kallikreins including EGF-binding protein types A, B, and C; NGF-γ chain, γ-renin, and prostate specific antigen (PSA); plasma kallikrein; mast cell proteases; myeloblastin (proteinase 3) (Wegener's autoantigen); plasminogen activators (urokinase-type, and tissue-type); and the trypsins I, II, III, and IV. These peptidases play key roles in coagulation, tumorigenesis, control of blood pressure, release of growth factors, and other roles. (http://www.babraham.co.uk/Merops/Merops.htm).

5. Threonine peptidases (T1) – (Prosite PDOC00326/PDOC00668)

Proteasomal subunits (T1A)

The proteasome is a multicatalytic threonine proteinase complex involved in ATP/ubiquitin dependent non-lysosomal proteolysis of cellular substrates. It is responsible for selective elimination of proteins with aberrant structures, as well as naturally occurring short-lived proteins related to metabolic regulation and cell-cycle progression (Momand *et al.*, 2000, *Gene* 242(1-2):15-29, Bochtler *et al.*, 1999, *Annu. Rev. Biophys Biomol Struct.*28:295-317). The proteasome inhibitor lactacystin reversibly inhibits proliferation of human endothelial cells, suggesting a role for proteasomes in angiogenesis (Kumeda, *et al.*, *Anticancer Res.* 1999 Sep-Oct;19(5B):3961-8). Another important function of the proteasome in higher vertebrates is to generate the peptides presented on MHC-class 1 molecules to circulating lymphocytes (Castelli *et al.*, 1997, *Int. J. Clin. Lab. Res.* 27(2):103-10). The proteasome has a sedimentation coefficient of 26S and is composed of a 20S catalytic core and a 22S regulatory complex. Eukaryotic 20S proteasomes have a molecular mass of 700 to 800 kD and consist of a set of over 15 kinds of polypeptides of 21 to 32 kD. All eukaryotic 20S proteasome subunits can be classified grossly into 2 subfamilies, α and β, by their high similarity with either the α or β subunits of the archaebacterium Thermoplasma acidophilum (Mayr *et al.*, 1999, *Biol. Chem.* 380(10):1183-92). Several of the components have been identified as threonine peptidases, suggesting that this class of peptidases plays a key

role in regulating metabolic pathways and cell-cycle progression, among other functions (Yorgin *et al.*, 2000, *J. Immunol.* 164(6):2915-23).

6. Peptidases of Unknown Catalytic Mechanism

The prenyl-protein specific protease responsible for post-translational processing of the Ras proto-oncogene and other prenylated proteins falls into this class. This class also includes several viral peptidases that may play a role in mammalian infection, including cardiovirus endopeptidase 2A (encephalomyocarditis virus) (Molla *et al.*, 1993, *J. Virol.* 67(8):4688-95), NS2-3 protease (hepatitis C virus) (Blight *et al.*, 1998, *Antivir. Ther.* 3(Suppl 3):71-81), endopeptidase (infectious pancreatic necrosis virus) (Lejal *et al.*, *J. Gen. Virol.*, 2000, 81:983-992), and the Npro endopeptidase (hog cholera virus) (Tratschin *et al.*, 1998, *J. Virol.* 72(9):7681-84).

Nucleic Acid Probes, Methods, and Kits for Detection of Proteases

A nucleic acid probe of the present invention may be used to probe an appropriate chromosomal or cDNA library by usual hybridization methods to obtain other nucleic acid molecules of the present invention. A chromosomal DNA or cDNA library may be prepared from appropriate cells according to recognized methods in the art (*cf.* "Molecular Cloning: A Laboratory Manual", second edition, Cold Spring Harbor Laboratory, Sambrook, Fritsch, & Maniatis, eds., 1989).

In the alternative, chemical synthesis can be carried out in order to obtain nucleic acid probes having nucleotide sequences which correspond to N-terminal and C-terminal portions of the amino acid sequence of the polypeptide of interest. The synthesized nucleic acid probes may be used as primers in a polymerase chain reaction (PCR) carried out in accordance with recognized PCR techniques, essentially according to PCR Protocols, "A Guide to Methods and Applications", Academic Press, Michael, *et al.*, eds., 1990, utilizing the appropriate chromosomal or cDNA library to obtain the fragment of the present invention.

One skilled in the art can readily design such probes based on the sequence disclosed herein using methods of computer alignment and sequence analysis known in the art ("Molecular Cloning: A Laboratory Manual", 1989, *supra*). The hybridization probes of the present invention can be labeled by standard labeling techniques such as with a radiolabel, enzyme label, fluorescent label, biotin-avidin label, chemiluminescence, and the like. After hybridization, the probes may be visualized using known methods.

The nucleic acid probes of the present invention include RNA, as well as DNA probes, such probes being generated using techniques known in the art. The nucleic acid probe may be immobilized on a solid support. Examples of such solid supports include, but are not limited to, plastics such as polycarbonate, complex carbohydrates such as agarose and sepharose, and acrylic resins, such as polyacrylamide and latex beads. Techniques for coupling nucleic acid probes to such solid supports are well known in the art.

The test samples suitable for nucleic acid probing methods of the present invention include, for example, cells or nucleic acid extracts of cells, or biological fluids. The samples used in the above-described methods will vary based on the assay format, the detection method and the nature of the tissues, cells or extracts to be assayed. Methods for preparing nucleic acid extracts of cells are well known in the art and can be readily adapted in order to obtain a sample which is compatible with the method utilized.

One method of detecting the presence of nucleic acids of the invention in a sample comprises (a) contacting said sample with the above-described nucleic acid probe under conditions such that hybridization occurs, and (b) detecting the presence of said probe bound to said nucleic acid molecule. One skilled in the art would select the nucleic acid probe according to techniques known in the art as described above. Samples to be tested include but should not be limited to RNA samples of human tissue.

A kit for detecting the presence of nucleic acids of the invention in a sample comprises at least one container means having disposed therein the above-described

nucleic acid probe. The kit may further comprise other containers comprising one or more of the following: wash reagents and reagents capable of detecting the presence of bound nucleic acid probe. Examples of detection reagents include, but are not limited to radiolabelled probes, enzymatic labeled probes (horseradish peroxidase, alkaline phosphatase), and affinity labeled probes (biotin, avidin, or steptavidin). Preferably, the kit further comprises instructions for use.

In detail, a compartmentalized kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers or strips of plastic or paper. Such containers allow the efficient transfer of reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such containers will include a container which will accept the test sample, a container which contains the probe or primers used in the assay, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, and the like), and containers which contain the reagents used to detect the hybridized probe, bound antibody, amplified product, or the like. One skilled in the art will readily recognize that the nucleic acid probes described in the present invention can readily be incorporated into one of the established kit formats which are well known in the art.

DNA Constructs Comprising a Protease Nucleic Acid Molecule and Cells Containing These Constructs.

The present invention also relates to a recombinant DNA molecule comprising, 5′ to 3′, a promoter effective to initiate transcription in a host cell and the above-described nucleic acid molecules. In addition, the present invention relates to a recombinant DNA molecule comprising a vector and an above-described nucleic acid molecule. The present invention also relates to a nucleic acid molecule comprising a transcriptional region functional in a cell, a sequence complementary to an RNA sequence encoding an amino acid sequence corresponding to the above-described polypeptide, and a transcriptional termination region functional in said

cell. The above-described molecules may be isolated and/or purified DNA molecules.

The present invention also relates to a cell or organism that contains an above-described nucleic acid molecule and thereby is capable of expressing a polypeptide. The polypeptide may be purified from cells which have been altered to express the polypeptide. A cell is said to be "altered to express a desired polypeptide" when the cell, through genetic manipulation, is made to produce a protein which it normally does not produce or which the cell normally produces at lower levels. One skilled in the art can readily adapt procedures for introducing and expressing either genomic, cDNA, or synthetic sequences into either eukaryotic or prokaryotic cells.

A nucleic acid molecule, such as DNA, is said to be "capable of expressing" a polypeptide if it contains nucleotide sequences which contain transcriptional and translational regulatory information and such sequences are "operably linked" to nucleotide sequences which encode the polypeptide. An operable linkage is a linkage in which the regulatory DNA sequences and the DNA sequence sought to be expressed are connected in such a way as to permit gene sequence expression. The precise nature of the regulatory regions needed for gene sequence expression may vary from organism to organism, but shall in general include a promoter region which, in prokaryotes, contains both the promoter (which directs the initiation of RNA transcription) as well as the DNA sequences which, when transcribed into RNA, will signal synthesis initiation. Such regions will normally include those 5′-non-coding sequences involved with initiation of transcription and translation, such as the TATA box, capping sequence, CAAT sequence, and the like.

If desired, the non-coding region 3′ to the sequence encoding a protease of the invention may be obtained by the above-described methods. This region may be retained for its transcriptional termination regulatory sequences, such as termination and polyadenylation. Thus, by retaining the 3′-region naturally contiguous to the DNA sequence encoding a protease of the invention, the transcriptional termination

signals may be provided. Where the transcriptional termination signals are not satisfactorily functional in the expression host cell, then a 3′ region functional in the host cell may be substituted.

Two DNA sequences (such as a promoter region sequence and a sequence encoding a protease of the invention) are said to be operably linked if the nature of the linkage between the two DNA sequences allows the protease sequence to be transcribed, i.e., where the linkage does not (1) result in the introduction of a frame-shift mutation, (2) interfere with the ability of the promoter region sequence to direct the transcription of a gene sequence encoding a protease of the invention, or (3) interfere with the ability of the gene sequence of a protease of the invention to be transcribed by the promoter region sequence. Thus, a promoter region would be operably linked to a DNA sequence if the promoter were capable of effecting transcription of that DNA sequence. Thus, to express a gene encoding a protease of the invention, transcriptional and translational signals recognized by an appropriate host are necessary.

The present invention encompasses the expression of a gene encoding a protease of the invention (or a functional derivative thereof) in either prokaryotic or eukaryotic cells. Prokaryotic hosts are, generally, very efficient and convenient for the production of recombinant proteins and are, therefore, one type of preferred expression system for proteases of the invention. Prokaryotes most frequently are represented by various strains of *E. coli*. However, other microbial strains may also be used, including other bacterial strains.

In prokaryotic systems, plasmid vectors that contain replication sites and control sequences derived from a species compatible with the host may be used. Examples of suitable plasmid vectors may include pBR322, pUC118, pUC119 and the like; suitable phage or bacteriophage vectors may include λgt10, λgt11 and the like; and suitable virus vectors may include pMAM-neo, pKRC and the like. Preferably, the selected vector of the present invention has the capacity to replicate in the selected host cell.

Recognized prokaryotic hosts include bacteria such as *E. coli, Bacillus, Streptomyces, Pseudomonas, Salmonella, Serratia*, and the like. However, under such conditions, the polypeptide will not be glycosylated. The prokaryotic host must be compatible with the replicon and control sequences in the expression plasmid.

To express a protease of the invention (or a functional derivative thereof) in a prokaryotic cell, it is necessary to operably link the sequence encoding the protease of the invention to a functional prokaryotic promoter. Such promoters may be either constitutive or, more preferably, regulatable (*i.e.*, inducible or derepressible). Examples of constitutive promoters include the *int* promoter of bacteriophage λ, the *bla* promoter of the β-lactamase gene sequence of pBR322, and the cat promoter of the chloramphenicol acetyl transferase gene sequence of pPR325, and the like. Examples of inducible prokaryotic promoters include the major right and left promoters of bacteriophage λ (P_L and P_R), the *trp*, *recA*, *λacZ*, *λacI*, and *gal* promoters of *E. coli*, the α-amylase (Ulmanen *et al.*, *J. Bacteriol.* 162:176-182, 1985) and the ς-28-specific promoters of *B. subtilis* (Gilman *et al.*, *Gene Sequence* 32:11-20, 1984), the promoters of the bacteriophages of *Bacillus* (Gryczan, in: The Molecular Biology of the Bacilli, Academic Press, Inc., NY, 1982), and *Streptomyces* promoters (Ward *et al.*, *Mol. Gen. Genet.* 203:468-478, 1986). Prokaryotic promoters are reviewed by Glick (*Ind. Microbiot.* 1:277-282, 1987), Cenatiempo (*Biochimie* 68:505-516, 1986), and Gottesman (*Ann. Rev. Genet.* 18:415-442, 1984).

Proper expression in a prokaryotic cell may also require the presence of a ribosome-binding site upstream of the gene sequence-encoding sequence. Such ribosome-binding sites are disclosed, for example, by Gold *et al.* (*Ann. Rev. Microbiol.* 35:365-404, 1981). The selection of control sequences, expression vectors, transformation methods, and the like, are dependent on the type of host cell used to express the gene. As used herein, "cell", "cell line", and "cell culture" may be used interchangeably and all such designations include progeny. Thus, the words

"transformants" or "transformed cells" include the primary subject cell and cultures derived therefrom, without regard to the number of transfers. It is also understood that all progeny may not be precisely identical in DNA content, due to deliberate or inadvertent mutations. However, as defined, mutant progeny have the same functionality as that of the originally transformed cell.

Host cells which may be used in the expression systems of the present invention are not strictly limited, provided that they are suitable for use in the expression of the protease polypeptide of interest. Suitable hosts may often include eukaryotic cells. Preferred eukaryotic hosts include, for example, yeast, fungi, insect cells, mammalian cells either *in vivo*, or in tissue culture. Mammalian cells which may be useful as hosts include HeLa cells, cells of fibroblast origin such as VERO or CHO-K1, or cells of lymphoid origin and their derivatives. Preferred mammalian host cells include SP2/0 and J558L, as well as neuroblastoma cell lines such as IMR 332, which may provide better capacities for correct post-translational processing.

In addition, plant cells are also available as hosts, and control sequences compatible with plant cells are available, such as the cauliflower mosaic virus 35S and 19S, and nopaline synthase promoter and polyadenylation signal sequences. Another preferred host is an insect cell, for example the *Drosophila* larvae. Using insect cells as hosts, the *Drosophila* alcohol dehydrogenase promoter can be used (Rubin, *Science* 240:1453-1459, 1988). Alternatively, baculovirus vectors can be engineered to express large amounts of proteases of the invention in insect cells (Jasny, *Science* 238:1653, 1987; Miller *et al.*, in: *Genetic Engineering*, Vol. 8, Plenum, Setlow *et al.*, eds., pp. 277-297, 1986).

Any of a series of yeast expression systems can be utilized which incorporate promoter and termination elements from the actively expressed sequences coding for glycolytic enzymes that are produced in large quantities when yeast are grown in mediums rich in glucose. Known glycolytic gene sequences can also provide very efficient transcriptional control signals. Yeast provides substantial advantages in that it can also carry out post-translational modifications. A number of recombinant

DNA strategies exist utilizing strong promoter sequences and high copy number plasmids which can be utilized for production of the desired proteins in yeast. Yeast recognizes leader sequences on cloned mammalian genes and secretes peptides bearing leader sequences (*i.e.*, pre-peptides). Several possible vector systems are available for the expression of proteases of the invention in a mammalian host.

A wide variety of transcriptional and translational regulatory sequences may be employed, depending upon the nature of the host. The transcriptional and translational regulatory signals may be derived from viral sources, such as adenovirus, bovine papilloma virus, cytomegalovirus, simian virus, or the like, where the regulatory signals are associated with a particular gene sequence which has a high level of expression. Alternatively, promoters from mammalian expression products, such as actin, collagen, myosin, and the like, may be employed. Transcriptional initiation regulatory signals may be selected which allow for repression or activation, so that expression of the gene sequences can be modulated. Of interest are regulatory signals which are temperature-sensitive so that by varying the temperature, expression can be repressed or initiated, or are subject to chemical (such as metabolite) regulation.

Expression of proteases of the invention in eukaryotic hosts requires the use of eukaryotic regulatory regions. Such regions will, in general, include a promoter region sufficient to direct the initiation of RNA synthesis. Preferred eukaryotic promoters include, for example, the promoter of the mouse metallothionein I gene sequence (Hamer *et al.*, *J. Mol. Appl. Gen.* 1:273-288, 1982); the TK promoter of Herpes virus (McKnight, *Cell* 31:355-365, 1982); the SV40 early promoter (Benoist *et al.*, *Nature* (London) 290:304-31, 1981); and the yeast gal4 gene sequence promoter (Johnston *et al.*, *Proc. Natl. Acad. Sci.* (USA) 79:6971-6975, 1982; Silver *et al.*, *Proc. Natl. Acad. Sci.* (USA) 81:5951-5955, 1984).

Translation of eukaryotic mRNA is initiated at the codon which encodes the first methionine. For this reason, it is preferable to ensure that the linkage between a eukaryotic promoter and a DNA sequence which encodes a protease of the invention (or a functional derivative thereof) does not contain any intervening codons which

are capable of encoding a methionine (*i.e.*, AUG). The presence of such codons results either in the formation of a fusion protein (if the AUG codon is in the same reading frame as the protease of the invention coding sequence) or a frame-shift mutation (if the AUG codon is not in the same reading frame as the protease of the invention coding sequence).

A nucleic acid molecule encoding a protease of the invention and an operably linked promoter may be introduced into a recipient prokaryotic or eukaryotic cell either as a nonreplicating DNA or RNA molecule, which may either be a linear molecule or, more preferably, a closed covalent circular molecule. Since such molecules are incapable of autonomous replication, the expression of the gene may occur through the transient expression of the introduced sequence. Alternatively, permanent expression may occur through the integration of the introduced DNA sequence into the host chromosome.

A vector may be employed which is capable of integrating the desired gene sequences into the host cell chromosome. Cells which have stably integrated the introduced DNA into their chromosomes can be selected by also introducing one or more markers which allow for selection of host cells which contain the expression vector. The marker may provide for prototrophy to an auxotrophic host, biocide resistance, *e.g.*, antibiotics, or heavy metals, such as copper, or the like. The selectable marker gene sequence can either be directly linked to the DNA gene sequences to be expressed, or introduced into the same cell by co-transfection. Additional elements may also be needed for optimal synthesis of mRNA. These elements may include splice signals, as well as transcription promoters, enhancers, and termination signals. cDNA expression vectors incorporating such elements include those described by Okayama (*Mol. Cell. Biol.* 3:280-289, 1983).

The introduced nucleic acid molecule can be incorporated into a plasmid or viral vector capable of autonomous replication in the recipient host. Any of a wide variety of vectors may be employed for this purpose. Factors of importance in selecting a particular plasmid or viral vector include: the ease with which recipient cells that contain the vector may be recognized and selected from those recipient

cells which do not contain the vector; the number of copies of the vector which are desired in a particular host; and whether it is desirable to be able to "shuttle" the vector between host cells of different species.

Preferred prokaryotic vectors include plasmids such as those capable of replication in *E. coli* (such as, for example, pBR322, ColEl, pSC101, pACYC 184, πVX; "Molecular Cloning: A Laboratory Manual", 1989, *supra*). Bacillus plasmids include pC194, pC221, pT127, and the like (Gryczan, In: The Molecular Biology of the Bacilli, Academic Press, NY, pp. 307-329, 1982). Suitable *Streptomyces* plasmids include p1J101 (Kendall *et al.*, *J. Bacteriol.* 169:4177-4183, 1987), and streptomyces bacteriophages such as ϕC31 (Chater *et al.*, In: Sixth International Symposium on Actinomycetales Biology, Akademiai Kaido, Budapest, Hungary, pp. 45-54, 1986). *Pseudomonas* plasmids are reviewed by John *et al.* (*Rev. Infect. Dis.* 8:693-704, 1986), and Izaki (*Jpn. J. Bacteriol.* 33:729-742, 1978).

Preferred eukaryotic plasmids include, for example, BPV, vaccinia, SV40, 2-micron circle, and the like, or their derivatives. Such plasmids are well known in the art (Botstein *et al.*, *Miami Wntr. Symp.* 19:265-274, 1982; Broach, In: The Molecular Biology of the Yeast Saccharomyces: Life Cycle and Inheritance, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, p. 445-470, 1981; Broach, *Cell* 28:203-204, 1982; Bollon *et al.*, *J. Clin. Hematol. Oncol.* 10:39-48, 1980; Maniatis, In: Cell Biology: A Comprehensive Treatise, Vol. 3, *Gene Sequence Expression*, Academic Press, NY, pp. 563-608, 1980).

Once the vector or nucleic acid molecule containing the construct(s) has been prepared for expression, the DNA construct(s) may be introduced into an appropriate host cell by any of a variety of suitable means, *i.e.*, transformation, transfection, conjugation, protoplast fusion, electroporation, particle gun technology, calcium phosphate-precipitation, direct microinjection, and the like. After the introduction of the vector, recipient cells are grown in a selective medium, which selects for the growth of vector-containing cells. Expression of the cloned gene(s) results in the production of a protease of the invention, or fragments thereof. This can take place in the transformed cells as such, or following the induction of these

cells to differentiate (for example, by administration of bromodeoxyuracil to neuroblastoma cells or the like). A variety of incubation conditions can be used to form the peptide of the present invention. The most preferred conditions are those which mimic physiological conditions.

Antibodies, Hybridomas, Methods of Use and Kits for Detection of Proteases

The present invention relates to an antibody having binding affinity to a protease of the invention. The protease polypeptide may have the amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or a functional derivative thereof, or at least 9 contiguous amino acids thereof (preferably, at least 20, 30, 35, or 40 contiguous amino acids thereof).

The present invention also relates to an antibody having specific binding affinity to a protease of the invention. Such an antibody may be isolated by comparing its binding affinity to a protease of the invention with its binding affinity to other polypeptides. Those which bind selectively to a protease of the invention would be chosen for use in methods requiring a distinction between a protease of the invention and other polypeptides. Such methods could include, but should not be limited to, the analysis of altered protease expression in tissue containing other polypeptides.

The proteases of the present invention can be used in a variety of procedures and methods, such as for the generation of antibodies, for use in identifying pharmaceutical compositions, and for studying DNA/protein interaction.

The proteases of the present invention can be used to produce antibodies or hybridomas. One skilled in the art will recognize that if an antibody is desired, such a peptide could be generated as described herein and used as an immunogen. The antibodies of the present invention include monoclonal and polyclonal antibodies, as well fragments of these antibodies, and humanized forms. Humanized forms of the antibodies of the present invention may be generated using one of the procedures known in the art such as chimerization or CDR grafting.

The present invention also relates to a hybridoma which produces the above-described monoclonal antibody, or binding fragment thereof. A hybridoma is an immortalized cell line which is capable of secreting a specific monoclonal antibody.

In general, techniques for preparing monoclonal antibodies and hybridomas are well known in the art (Campbell, Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology, Elsevier Science Publishers, Amsterdam, The Netherlands, 1984; St. Groth *et al.*, *J. Immunol. Methods* 35:1-21, 1980). Any animal (mouse, rabbit, and the like) which is known to produce antibodies can be immunized with the selected polypeptide. Methods for immunization are well known in the art. Such methods include subcutaneous or intraperitoneal injection of the polypeptide. One skilled in the art will recognize that the amount of polypeptide used for immunization will vary based on the animal which is immunized, the antigenicity of the polypeptide and the site of injection.

The polypeptide may be modified or administered in an adjuvant in order to increase the peptide antigenicity. Methods of increasing the antigenicity of a polypeptide are well known in the art. Such procedures include coupling the antigen with a heterologous protein (such as globulin or β-galactosidase) or through the inclusion of an adjuvant during immunization.

For monoclonal antibodies, spleen cells from the immunized animals are removed, fused with myeloma cells, such as SP2/0-Agl4 myeloma cells, and allowed to become monoclonal antibody producing hybridoma cells. Any one of a number of methods well known in the art can be used to identify the hybridoma cell which produces an antibody with the desired characteristics. These include

screening the hybridomas with an ELISA assay, western blot analysis, or radioimmunoassay (Lutz *et al.*, *Exp. Cell Res.* 175:109-124, 1988). Hybridomas secreting the desired antibodies are cloned and the class and subclass are determined using procedures known in the art (Campbell, "Monoclonal Antibody Technology: Laboratory Techniques in Biochemistry and Molecular Biology", *supra*, 1984).

For polyclonal antibodies, antibody-containing antisera is isolated from the immunized animal and is screened for the presence of antibodies with the desired specificity using one of the above-described procedures. The above-described antibodies may be detectably labeled. Antibodies can be detectably labeled through the use of radioisotopes, affinity labels (such as biotin, avidin, and the like), enzymatic labels (such as horseradish peroxidase, alkaline phosphatase, and the like) fluorescent labels (such as FITC or rhodamine, and the like), paramagnetic atoms, and the like. Procedures for accomplishing such labeling are well-known in the art, for example, *see* Stemberger *et al.*, J. *Histochem. Cytochem.* 18:315, 1970; Bayer *et al.*, *Meth. Enzym.* 62:308, 1979; Engval *et al.*, *Immunol.* 109:129, 1972; Goding, *J. Immunol. Meth.* 13:215, 1976. The antibodies of the present invention may be indirectly labelled by the use of secondary labelled antibodies, such as labelled anti-rabbit antibodies.The labeled antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays to identify cells or tissues which express a specific peptide.

The above-described antibodies may also be immobilized on a solid support. Examples of such solid supports include plastics such as polycarbonate, complex carbohydrates such as agarose and sepharose, acrylic resins such as polyacrylamide and latex beads. Techniques for coupling antibodies to such solid supports are well known in the art (Weir *et al.*, "Handbook of Experimental Immunology" 4th Ed., Blackwell Scientific Publications, Oxford, England, Chapter 10, 1986; Jacoby *et al.*, *Meth. Enzym.* 34, Academic Press, N.Y., 1974). The immobilized antibodies of the present invention can be used for *in vitro*, *in vivo*, and *in situ* assays as well as in immunochromotography.

Furthermore, one skilled in the art can readily adapt currently available procedures, as well as the techniques, methods and kits disclosed herein with regard to antibodies, to generate peptides capable of binding to a specific peptide sequence in order to generate rationally designed antipeptide peptides (Hurby *et al.*, "Application of Synthetic Peptides: Antisense Peptides", In Synthetic Peptides, A User's Guide, W.H. Freeman, NY, pp. 289-307, 1992; Kaspczak *et al.*, *Biochemistry* 28:9230-9238, 1989).

Anti-peptide peptides can be generated by replacing the basic amino acid residues found in the peptide sequences of the proteases of the invention with acidic residues, while maintaining hydrophobic and uncharged polar groups. For example, lysine, arginine, and/or histidine residues are replaced with aspartic acid or glutamic acid and glutamic acid residues are replaced by lysine, arginine or histidine.

The present invention also encompasses a method of detecting a protease polypeptide in a sample, comprising: (a) contacting the sample with an above-described antibody, under conditions such that immunocomplexes form, and (b) detecting the presence of said antibody bound to the polypeptide. In detail, the methods comprise incubating a test sample with one or more of the antibodies of the present invention and assaying whether the antibody binds to the test sample. Altered levels of a protease of the invention in a sample as compared to normal levels may indicate disease.

Conditions for incubating an antibody with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the antibody used in the assay. One skilled in the art will recognize that any one of the commonly available immunological assay formats (such as radioimmunoassays, enzyme-linked immunosorbent assays, diffusion-based Ouchterlony, or rocket immunofluorescent assays) can readily be adapted to employ the antibodies of the present invention. Examples of such assays can be found in Chard ("An Introduction to Radioimmunoassay and Related Techniques" Elsevier Science Publishers, Amsterdam, The Netherlands, 1986), Bullock *et al.* ("Techniques in Immunocytochemistry," Academic Press, Orlando,

FL Vol. 1, 1982; Vol. 2, 1983; Vol. 3, 1985), Tijssen ("Practice and Theory of Enzyme Immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology," Elsevier Science Publishers, Amsterdam, The Netherlands, 1985).

The immunological assay test samples of the present invention include cells, protein or membrane extracts of cells, or biological fluids such as blood, serum, plasma, or urine. The test samples used in the above-described method will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods for preparing protein extracts or membrane extracts of cells are well known in the art and can readily be adapted in order to obtain a sample which is testable with the system utilized.

A kit contains all the necessary reagents to carry out the previously described methods of detection. The kit may comprise: (i) a first container means containing an above-described antibody, and (ii) second container means containing a conjugate comprising a binding partner of the antibody and a label. In another preferred embodiment, the kit further comprises one or more other containers comprising one or more of the following: wash reagents and reagents capable of detecting the presence of bound antibodies.

Examples of detection reagents include, but are not limited to, labeled secondary antibodies, or in the alternative, if the primary antibody is labeled, the chromophoric, enzymatic, or antibody binding reagents which are capable of reacting with the labeled antibody. The compartmentalized kit may be as described above for nucleic acid probe kits. One skilled in the art will readily recognize that the antibodies described in the present invention can readily be incorporated into one of the established kit formats which are well known in the art.

Isolation of Compounds Which Interact with Proteases

The present invention also relates to a method of detecting a compound capable of binding to a protease of the invention comprising incubating the compound with a protease of the invention and detecting the presence of the

compound bound to the protease. The compound may be present within a complex mixture, for example, serum, body fluid, or cell extracts.

The present invention also relates to a method of detecting an agonist or antagonist of protease activity or protease binding partner activity comprising incubating cells that produce a protease of the invention in the presence of a compound and detecting changes in the level of protease activity or protease binding partner activity. The compounds thus identified would produce a change in activity indicative of the presence of the compound. The compound may be present within a complex mixture, for example, serum, body fluid, or cell extracts. Once the compound is identified it can be isolated using techniques well known in the art.

The present invention also encompasses a method modulating protease associated activity in a mammal comprising administering to said mammal an agonist or antagonist to a protease of the invention in an amount sufficient to effect said modulation. A method of treating diseases in a mammal with an agonist or antagonist of the activity of one of the proteases of the invention comprising administering the agonist or antagonist to a mammal in an amount sufficient to agonize or antagonize protease-associated functions is also encompassed in the present application.

In an effort to discover novel treatments for diseases, biomedical researchers and chemists have designed, synthesized, and tested molecules that inhibit the function of proteases. Some small organic molecules form a class of compounds that modulate the function of protein proteases.

Examples of molecules that have been reported to inhibit the function of protein proteases include, but are not limited to, phenylmethylsulfonyl fluoride (PMSF), diisopropylfluorophosphate (DFP) (chapter 3, Barrett *et al.*, Handbook of Proteolytic Enzymes, 1998, Academic Press, San Diego), 3,4-dichloroisocoumarin (DCI) (*Id.*, chapter 16), serpins (*Id.*, chapter 37), E-64 (*trans*-epoxysuccinyl L-leucylamido-(4-guanidino) butane) (*Id.*, chapter 188), peptidyl-diazomethanes, peptidyl-*O*-acyl-hydroxamates, epoxysuccinyl-peptides (*Id.*, chapter 210), DAN, EPNP (1,2-epoxy-3(p-nitrophenoxy)propane) (*Id.*, chapter 298), thiorphan (dl-3-

Mercapto-2-benzylpropanoyl-glycine) (*Id.*, chapter 362), CGS 26303, PD 069185 (*Id.*, chapter 363), and COT989-00 (N-4-hydroxy-N1-[1-(s)-(4-aminosulfonyl)phenylethyl-aminocarboxyl-2-cyclohexylethyl)-2R-[4-methyl)phenylpropyl]succinamide) (*Id.*, chapter 401). Other protease inhibitors include, but are not limited to, aprotinin, amastatin, antipain, calcineurin autoinhibitory fragment, and histatin 5 (*Id.*). Preferably, these inhibitors will have molecular weights from 100 to 200 daltons, from 200 to 300 daltons, from 300 to 400 daltons, from 400 to 600 daltons, from 600 to 1000 daltons, from 1000 to 2000 daltons, from 2000 to 4000 daltons, and from 4000 to 8000 daltons.

Compounds that can traverse cell membranes and are resistant to acid hydrolysis are potentially advantageous as therapeutics as they can become highly bioavailable after being administered orally to patients. However, many of these protease inhibitors only weakly inhibit the function of proteases. In addition, many inhibit a variety of proteases and will therefore cause multiple side-effects as therapeutics for diseases.

Transgenic Animals.

A variety of methods are available for the production of transgenic animals associated with this invention. DNA can be injected into the pronucleus of a fertilized egg before fusion of the male and female pronuclei, or injected into the nucleus of an embryonic cell (*e.g.*, the nucleus of a two-cell embryo) following the initiation of cell division (Brinster *et al.*, *Proc. Nat. Acad. Sci.* USA 82:4438-4442, 1985). Embryos can be infected with viruses, especially retroviruses, modified to carry inorganic-ion receptor nucleotide sequences of the invention.

Pluripotent stem cells derived from the inner cell mass of the embryo and stabilized in culture can be manipulated in culture to incorporate nucleotide sequences of the invention. A transgenic animal can be produced from such cells through implantation into a blastocyst that is implanted into a foster mother and allowed to come to term. Animals suitable for transgenic experiments can be

obtained from standard commercial sources such as Charles River (Wilmington, MA), Taconic (Germantown, NY), Harlan Sprague Dawley (Indianapolis, IN), etc.

The procedures for manipulation of the rodent embryo and for microinjection of DNA into the pronucleus of the zygote are well known to those of ordinary skill in the art (Hogan *et al.*, *supra*). Microinjection procedures for fish, amphibian eggs and birds are detailed in Houdebine and Chourrout (*Experientia* 47:897-905, 1991). Other procedures for introduction of DNA into tissues of animals are described in U.S. Patent No. 4,945,050 (Sanford *et al.*, July 30, 1990).

By way of example only, to prepare a transgenic mouse, female mice are induced to superovulate. Females are placed with males, and the mated females are sacrificed by CO_2 asphyxiation or cervical dislocation and embryos are recovered from excised oviducts. Surrounding cumulus cells are removed. Pronuclear embryos are then washed and stored until the time of injection. Randomly cycling adult female mice are paired with vasectomized males. Recipient females are mated at the same time as donor females. Embryos then are transferred surgically. The procedure for generating transgenic rats is similar to that of mice (Hammer *et al.*, *Cell* 63:1099-1112, 1990).

Methods for the culturing of embryonic stem (ES) cells and the subsequent production of transgenic animals by the introduction of DNA into ES cells using methods such as electroporation, calcium phosphate/DNA precipitation and direct injection also are well known to those of ordinary skill in the art (Teratocarcinomas and Embryonic Stem Cells, A Practical Approach, E.J. Robertson, ed., IRL Press, 1987).

In cases involving random gene integration, a clone containing the sequence(s) of the invention is co-transfected with a gene encoding resistance. Alternatively, the gene encoding neomycin resistance is physically linked to the sequence(s) of the invention. Transfection and isolation of desired clones are carried out by any one of several methods well known to those of ordinary skill in the art (E.J. Robertson, *supra*).

DNA molecules introduced into ES cells can also be integrated into the chromosome through the process of homologous recombina-tion (Capecchi, *Science* 244:1288-1292, 1989). Methods for positive selection of the recombination event (*i.e.*, neo resistance) and dual positive-negative selection (*i.e.*, neo resistance and gancyclovir resistance) and the subsequent identification of the desired clones by PCR have been described by Capecchi, *supra* and Joyner *et al.* (*Nature* 338:153-156, 1989), the teachings of which are incorporated herein in their entirety including any drawings. The final phase of the procedure is to inject targeted ES cells into blastocysts and to transfer the blastocysts into pseudopregnant females. The resulting chimeric animals are bred and the offspring are analyzed by Southern blotting to identify individuals that carry the transgene. Procedures for the production of non-rodent mammals and other animals have been discussed by others (Houdebine and Chourrout, *supra*; Pursel *et al.*, *Science* 244:1281-1288, 1989; and Simms *et al.*, *Bio/Technology* 6:179-183, 1988).

Thus, the invention provides transgenic, nonhuman mammals containing a transgene encoding a protease of the invention or a gene affecting the expression of the protease. Such transgenic nonhuman mammals are particularly useful as an *in vivo* test system for studying the effects of introduction of a protease, or regulating the expression of a protease (*i.e.*, through the introduction of additional genes, antisense nucleic acids, or ribozymes).

A "transgenic animal" is an animal having cells that contain DNA which has been artificially inserted into a cell, which DNA becomes part of the genome of the animal which develops from that cell. Preferred transgenic animals are primates, mice, rats, cows, pigs, horses, goats, sheep, dogs and cats. The transgenic DNA may encode human proteases. Native expression in an animal may be reduced by providing an amount of antisense RNA or DNA effective to reduce expression of the receptor.

Gene Therapy

Proteases or their genetic sequences will also be useful in gene therapy (reviewed in Miller, *Nature* 357:455-460, 1992). Miller states that advances have resulted in practical approaches to human gene therapy that have demonstrated positive initial results. The basic science of gene therapy is described in Mulligan (*Science* 260:926-931, 1993).

In one preferred embodiment, an expression vector containing a protease coding sequence is inserted into cells, the cells are grown *in vitro* and then infused in large numbers into patients. In another preferred embodiment, a DNA segment containing a promoter of choice (for example a strong promoter) is transferred into cells containing an endogenous gene encoding proteases of the invention in such a manner that the promoter segment enhances expression of the endogenous protease gene (for example, the promoter segment is transferred to the cell such that it becomes directly linked to the endogenous protease gene).

The gene therapy may involve the use of an adenovirus containing protease cDNA targeted to a tumor, systemic protease increase by implantation of engineered cells, injection with protease-encoding virus, or injection of naked protease DNA into appropriate tissues.

Target cell populations may be modified by introducing altered forms of one or more components of the protein complexes in order to modulate the activity of such complexes. For example, by reducing or inhibiting a complex component activity within target cells, an abnormal signal transduction event(s) leading to a condition may be decreased, inhibited, or reversed. Deletion or missense mutants of a component, that retain the ability to interact with other components of the protein complexes but cannot function in signal transduction, may be used to inhibit an abnormal, deleterious signal transduction event.

Expression vectors derived from viruses such as retroviruses, vaccinia virus, adenovirus, adeno-associated virus, herpes viruses, several RNA viruses, or bovine papilloma virus, may be used for delivery of nucleotide sequences (*e.g.*, cDNA) encod-ing recombinant protease of the invention protein into the targeted cell

population (*e.g.*, tumor cells). Methods which are well known to those skilled in the art can be used to construct recombinant viral vectors containing coding sequences (Maniatis *et al.*, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, N.Y., 1989; Ausubel *et al.*, Current Protocols in Molecular Biology, Greene Publishing Associates and Wiley Interscience, N.Y., 1989). Alternatively, recombinant nucleic acid molecules encoding protein sequences can be used as naked DNA or in a reconstituted system *e.g.*, liposomes or other lipid systems for delivery to target cells (*e.g.*, Felgner *et al.*, *Nature* 337:387-8, 1989). Several other methods for the direct transfer of plasmid DNA into cells exist for use in human gene therapy and involve targeting the DNA to receptors on cells by complexing the plasmid DNA to proteins (Miller, *supra*).

In its simplest form, gene transfer can be performed by simply injecting minute amounts of DNA into the nucleus of a cell, through a process of microinjection (Capecchi, *Cell* 22:479-88, 1980). Once recombinant genes are introduced into a cell, they can be recognized by the cell's normal mechanisms for transcription and translation, and a gene product will be expressed. Other methods have also been attempted for introducing DNA into larger numbers of cells. These methods include: transfection, wherein DNA is precipitated with calcium phosphate and taken into cells by pinocytosis (Chen *et al.*, *Mol. Cell Biol.* 7:2745-52, 1987); electroporation, wherein cells are exposed to large voltage pulses to introduce holes into the membrane (Chu *et al.*, *Nucleic Acids Res.* 15:1311-26, 1987); lipofection/liposome fusion, wherein DNA is packaged into lipophilic vesicles which fuse with a target cell (Felgner *et al.*, *Proc. Natl. Acad. Sci.* USA. 84:7413-7417, 1987); and particle bombardment using DNA bound to small projectiles (Yang *et al.*, *Proc. Natl. Acad. Sci.* 87:9568-9572, 1990). Another method for introducing DNA into cells is to couple the DNA to chemically modified proteins.

It has also been shown that adenovirus proteins are capable of destabilizing endosomes and enhancing the uptake of DNA into cells. The admixture of adenovirus to solutions containing DNA complexes, or the binding of DNA to polylysine covalently attached to adenovirus using protein crosslinking agents

substantially improves the uptake and expression of the recombinant gene (Curiel *et al.*, *Am. J. Respir. Cell. Mol. Biol.*, 6:247-52, 1992).

As used herein "gene transfer" means the process of introducing a foreign nucleic acid molecule into a cell. Gene transfer is commonly performed to enable the expression of a particular product encoded by the gene. The product may include a protein, polypeptide, anti-sense DNA or RNA, or enzymatically active RNA. Gene transfer can be performed in cultured cells or by direct administration into animals. Generally gene transfer involves the process of nucleic acid contact with a target cell by non-specific or receptor mediated interactions, uptake of nucleic acid into the cell through the membrane or by endocytosis, and release of nucleic acid into the cyto-plasm from the plasma membrane or endosome. Expression may require, in addition, movement of the nucleic acid into the nucleus of the cell and binding to appropriate nuclear factors for transcription.

As used herein "gene therapy" is a form of gene transfer and is included within the definition of gene transfer as used herein and specifically refers to gene transfer to express a therapeutic product from a cell *in vivo* or *in vitro*. Gene transfer can be performed *ex vivo* on cells which are then transplanted into a patient, or can be performed by direct administration of the nucleic acid or nucleic acid-protein complex into the patient.

In another preferred embodiment, a vector having nucleic acid sequences encoding a protease polypeptide is provided in which the nucleic acid sequence is expressed only in specific tissue. Methods of achieving tissue-specific gene expression are set forth in International Publication No. WO 93/09236, filed November 3, 1992 and published May 13, 1993.

In all of the preceding vectors set forth above, a further aspect of the invention is that the nucleic acid sequence contained in the vector may include additions, deletions or modifications to some or all of the sequence of the nucleic acid, as defined above.

Expression, including over-expression, of a protease polypeptide of the invention can be inhibited by administration of an antisense molecule that binds to and

inhibits expression of the mRNA encoding the polypeptide. Alternatively, expression can be inhibited in an analogous manner using a ribozyme that cleaves the mRNA. General methods of using antisense and ribozyme technology to control gene expression, or of gene therapy methods for expression of an exogenous gene in this manner are well known in the art. Each of these methods utilizes a system, such as a vector, encoding either an antisense or ribozyme transcript of a protease polypeptide of the invention.

The term "***ribozyme***" refers to an RNA structure of one or more RNAs having catalytic properties. Ribozymes generally exhibit endonuclease, ligase or polymerase activity. Ribozymes are structural RNA molecules which mediate a number of RNA self-cleavage reactions. Various types of trans-acting ribozymes, including "hammerhead" and "hairpin" types, which have different secondary structures, have been identified. A variety of ribozymes have been characterized. See, for example, U.S. Pat. Nos. 5,246,921, 5,225,347, 5,225,337 and 5,149,796. Mixed ribozymes comprising deoxyribo and ribooligonucleotides with catalytic activity have been described. Perreault, *et al.*, *Nature*, 344:565-567 (1990).

As used herein, "antisense" refers of nucleic acid molecules or their derivatives which specifically hybridize, *e.g.*, bind, under cellular conditions, with the genomic DNA and/or cellular mRNA encoding a protease polypeptide of the invention, so as to inhibit expression of that protein, for example, by inhibiting transcription and/or translation. The binding may be by conventional base pair complementarity, or, for example, in the case of binding to DNA duplexes, through specific interactions in the major groove of the double helix.

In one aspect, the antisense construct is an nucleic acid which is generated *ex vivo* and that, when introduced into the cell, can inhibit gene expression by, without limitation, hybridizing with the mRNA and/or genomic sequences of a protease polynucleotide of the invention.

Antisense approaches can involve the design of oligonucleotides (either DNA or RNA) that are complementary to protease polypeptide mRNA and are based on the protease polynucleotides of the invention, including SEQ ID NO:1,

SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35. The antisense oligonucleotides will bind to the protease polypeptide mRNA transcripts and prevent translation.

Although absolute complementarity is preferred, it is not required. A sequence "complementary" to a portion of an RNA, as referred to herein, means a sequence having sufficient complementarity to be able to hybridize with the RNA, forming a stable duplex; in the case of double-stranded antisense nucleic acids, a single strand of the duplex DNA may thus be tested, or triplex formation may be assayed. The ability to hybridize will depend on both the degree of complementarity and the length of the antisense nucleic acid. Generally, the longer the hybridizing nucleic acid, the more base mismatches with an RNA it may contain and still form a stable duplex (or triplex, as the case may be). One skilled in the art can ascertain a tolerable degree of mismatch by use of standard procedures to determine the melting point of the hybridized complex.

In general, oligonucleotides that are complementary to the 5' end of the message, *e.g.*, the 5' untranslated sequence up to and including the AUG initiation codon, should work most efficiently at inhibiting translation. However, sequences complementary to the 3' untranslated sequences of mRNAs have been shown to be effective at inhibiting translation of mRNAs as well. (Wagner, R. (1994) Nature 372:333). Antisense oligonucleotides complementary to mRNA coding regions are less efficient inhibitors of translation but could be used in accordance with the invention. Whether designed to hybridize to the 5', 3' or coding region of the protease polypeptide mRNA, antisense nucleic acids should be at least six nucleotides in length, and are preferably less than about 100 and more preferably

less than about 50 or 30 nucleotides in length. Typically they should be between 10 and 25 nucleotides in length. Such principles will inform the practitioner in selecting the appropriate oligonucleotides In preferred embodiments, the antisense sequence is selected from an oligonucleotide sequence that comprises, consists of, or consists essentially of about 10-30, and more preferably 15-25, contiguous nucleotide bases of a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35 or domains thereof.

In another preferred embodiment, the invention includes an isolated, enriched or purified nucleic acid molecule comprising, consisting of or consisting essentially of about 10-30, and more preferably 15-25 contiguous nucleotide bases of a nucleic acid sequence that encodes a polypeptide that selected from the group consisting of SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

Using the sequences of the present invention, antisense oligonucleotides can be designed. Such antisense oligonucleotides would be administered to cells expressing the target protease and the levels of the target RNA or protein with that of an internal control RNA or protein would be compared. Results obtained using

the antisense oligonucleotide would also be compared with those obtained using a suitable control oligonucleotide. A preferred control oligonucleotide is an oligonucleotide of approximately the same length as the test oligonucleotide. Those antisense oligonucleotides resulting in a reduction in levels of target RNA or protein would be selected.

The oligonucleotides can be DNA or RNA or chimeric mixtures or derivatives or modified versions thereof, single-stranded or double-stranded. The oligonucleotide can be modified at the base moiety, sugar moiety, or phosphate backbone, for example, to improve stability of the molecule, hybridization, etc. The oligonucleotide may include other appended groups such as peptides (*e.g.*, for targeting host cell receptors *in vivo*), or agents facilitating transport across the cell membrane (*see, e.g.*, Letsinger *et al.* (1989) *Proc. Natl. Acad. Sci.* U.S.A. 86:6553-6556; Lemaitre *et al.* (1987) *Proc. Natl. Acad. Sci.* USA 84:648-652; PCT Publication No. WO 88/09810, published Dec. 15, 1988) or the blood-brain barrier (see, *e.g.*, PCT Publication No. WO 89/10134, published Apr. 25, 1988), hybridization-triggered cleavage agents. (*See, e.g.*, Krol *et al.* (1988) BioTechniques 6:958-976) or intercalating agents. (*See*, e.g, Zon (1988) *Pharm. Res.* 5:539-549). To this end, the oligonucleotide may be conjugated to another molecule, *e.g.*, a peptide, hybridization triggered cross-linking agent, transport agent, hybridization-triggered cleavage agent, etc.

The antisense oligonucleotide may comprise at least one modified base moiety which is selected from moieties such as 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, 4-acetylcytosine, and 5-(carboxyhydroxyethyl) uracil. The antisense oligonucleotide may also comprise at least one modified sugar moiety selected from the group including but not limited to arabinose, 2-fluoroarabinose, xylulose, and hexose.

In yet another embodiment, the antisense oligonucleotide comprises at least one modified phosphate backbone selected from the group consisting of a phosphorothioate, a phosphorodithioate, a phosphoramidothioate, a phosphoramidate, a phosphordiamidate, a methylphosphonate, an alkyl

phosphotriester, and a formacetal or analog thereof. (*see also* U.S. Pat. Nos. 5,176,996; 5,264,564; and 5,256,775)

In yet a further embodiment, the antisense oligonucleotide is an α-anomeric oligonucleotide. An α-anomeric oligonucleotide forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual β-units, the strands run parallel to each other (Gautier *et al.* (1987) *Nucl. Acids Res.* 15:6625-6641). The oligonucleotide is a 2'-0-methylribonucleotide (Inoue *et al.* (1987) *Nucl. Acids Res.* 15:6131-6148), or a chimeric RNA-DNA analogue (Inoue *et al.* (1987) FEBS Lett. 215:327-330).

Also suitable are peptidyl nucleic acids, which are polypeptides such as polyserine, polythreonine, etc. including copolymers containing various amino acids, which are substituted at side-chain positions with nucleic acids (T,A,G,C,U). Chains of such polymers are able to hybridize through complementary bases in the same manner as natural DNA/RNA.. Alternatively, an antisense construct of the present invention can be delivered, for example, as an expression plasmid or vector that, when transcribed in the cell, produces RNA complementary to at least a unique portion of the cellular mRNA which encodes a protease polypeptide of the invention.

While antisense nucleotides complementary to the protease polypeptide coding region sequence can be used, those complementary to the transcribed untranslated region are most preferred.

In another preferred embodiment, a method of gene replacement is set forth. "Gene replacement" as used herein means supplying a nucleic acid sequence which is capable of being expressed *in vivo* in an animal and thereby providing or augmenting the function of an endogenous gene which is missing or defective in the animal.

Pharmaceutical Formulations And Routes Of Administration

The compounds described herein, including protease polypeptides of the invention, antisense molecules, ribozymes, and any other compound that modulates the activity of a protease polypeptide of the invention, can be administered to a human patient *per se*, or in pharmaceutical compositions where it is mixed with other active ingredients, as in combination therapy, or suitable carriers or excipient(s). Techniques for formulation and administration of the compounds of the instant application may be found in "Remington's Pharmaceutical Sciences," Mack Publishing Co., Easton, PA, latest edition.

A. Routes Of Administration

Suitable routes of administration may, for example, include oral, rectal, transmucosal, or intestinal administration; parenteral delivery, including intramuscular, subcutaneous, intravenous, intramedullary injections, as well as intrathecal, direct intraventricular, intraperitoneal, intranasal, or intraocular injections.

Alternately, one may administer the compound in a local rather than systemic manner, for example, via injection of the compound directly into a solid tumor, often in a depot or sustained release formulation.

Furthermore, one may administer the drug in a targeted drug delivery system, for example, in a liposome coated with tumor-specific antibody. The liposomes will be targeted to and taken up selectively by the tumor.

B. Composition/Formulation

The pharmaceutical compositions of the present invention may be manufactured in a manner that is itself known, *e.g.*, by means of conventional mixing, dissolving, granulating, dragee-making, levigating, emulsifying, encapsulating, entrapping or lyophilizing processes.

Pharmaceutical compositions for use in accordance with the present invention thus may be formulated in conventional manner using one or more physiologically acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used

pharmaceutically. Proper formulation is dependent upon the route of administration chosen.

For injection, the agents of the invention may be formulated in aqueous solutions, preferably in physiologically compatible buffers such as Hanks's solution, Ringer's solution, or physiological saline buffer. For transmucosal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art.

For oral administration, the compounds can be formulated readily by combining the active compounds with pharmaceutically acceptable carriers well known in the art. Such carriers enable the compounds of the invention to be formulated as tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, suspensions and the like, for oral ingestion by a patient to be treated. Suitable carriers include excipients such as, fillers such as sugars, including lactose, sucrose, mannitol, or sorbitol; cellulose preparations such as, for example, maize starch, wheat starch, rice starch, potato starch, gelatin, gum tragacanth, methyl cellulose, hydroxypropylmethyl- cellulose, sodium carboxymethylcellulose, and/or polyvinylpyrrolidone (PVP). If desired, disintegrating agents may be added, such as the cross-linked polyvinyl pyrrolidone, agar, or alginic acid or a salt thereof such as sodium alginate.

Dragee cores are provided with suitable coatings. For this purpose, concentrated sugar solutions may be used, which may optionally contain gum arabic, talc, polyvinyl pyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures. Dyestuffs or pigments may be added to the tablets or dragee coatings for identification or to characterize different combinations of active compound doses.

Pharmaceutical preparations which can be used orally include push-fit capsules made of gelatin, as well as soft, sealed capsules made of gelatin and a plasticizer, such as glycerol or sorbitol. The push-fit capsules can contain the active ingredients in admixture with filler such as lactose, binders such as starches, and/or lubricants such as talc or magnesium stearate and, optionally, stabilizers. In soft

capsules, the active compounds may be dissolved or suspended in suitable liquids, such as fatty oils, liquid paraffin, or liquid polyethylene glycols. In addition, stabilizers may be added. All formulations for oral administration should be in dosages suitable for such administration.

For buccal administration, the compositions may take the form of tablets or lozenges formulated in conventional manner.

For administration by inhalation, the compounds for use according to the present invention are conveniently delivered in the form of an aerosol spray presentation from pressurized packs or a nebuliser, with the use of a suitable propellant, *e.g.*, dichlorodifluoromethane, trichlorofluoromethane, dichlorotetrafluoroethane, carbon dioxide or other suitable gas. In the case of a pressurized aerosol the dosage unit may be determined by providing a valve to deliver a metered amount. Capsules and cartridges of *e.g.* gelatin for use in an inhaler or insufflator may be formulated containing a powder mix of the compound and a suitable powder base such as lactose or starch.

The compounds may be formulated for parenteral administration by injection, *e.g.*, by bolus injection or continuous infusion. Formulations for injection may be presented in unit dosage form, *e.g.*, in ampoules or in multi-dose containers, with an added preservative. The compositions may take such forms as suspensions, solutions or emulsions in oily or aqueous vehicles, and may contain formulatory agents such as suspending, stabilizing and/or dispersing agents.

Pharmaceutical formulations for parenteral administration include aqueous solutions of the active compounds in water-soluble form. Additionally, suspensions of the active compounds may be prepared as appropriate oily injection suspensions. Suitable lipophilic solvents or vehicles include fatty oils such as sesame oil, or synthetic fatty acid esters, such as ethyl oleate or triglycerides, or liposomes. Aqueous injection suspensions may contain substances which increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Optionally, the suspension may also contain suitable stabilizers or agents which

increase the solubility of the compounds to allow for the preparation of highly concentrated solutions.

Alternatively, the active ingredient may be in powder form for constitution with a suitable vehicle, *e.g.*, sterile pyrogen-free water, before use.

The compounds may also be formulated in rectal compositions such as suppositories or retention enemas, *e.g.*, containing conventional suppository bases such as cocoa butter or other glycerides.

In addition to the formulations described previously, the compounds may also be formulated as a depot preparation. Such long acting formulations may be administered by implantation (for example subcutaneously or intramuscularly) or by intramuscular injection. Thus, for example, the compounds may be formulated with suitable polymeric or hydrophobic materials (for example as an emulsion in an acceptable oil) or ion exchange resins, or as sparingly soluble derivatives, for example, as a sparingly soluble salt.

A pharmaceutical carrier for the hydrophobic compounds of the invention is a cosolvent system comprising benzyl alcohol, a nonpolar surfactant, a water-miscible organic polymer, and an aqueous phase. The cosolvent system may be the VPD co-solvent system. VPD is a solution of 3% w/v benzyl alcohol, 8% w/v of the nonpolar surfactant polysorbate 80, and 65% w/v polyethylene glycol 300, made up to volume in absolute ethanol. The VPD co-solvent system (VPD:D5W) consists of VPD diluted 1:1 with a 5% dextrose in water solution. This co-solvent system dissolves hydrophobic compounds well, and itself produces low toxicity upon systemic administration. Naturally, the proportions of a co-solvent system may be varied considerably without destroying its solubility and toxicity characteristics. Furthermore, the identity of the co-solvent components may be varied: for example, other low-toxicity nonpolar surfactants may be used instead of polysorbate 80; the fraction size of polyethylene glycol may be varied; other biocompatible polymers may replace polyethylene glycol, *e.g.* polyvinyl pyrrolidone; and other sugars or polysaccharides may substitute for dextrose.

Alternatively, other delivery systems for hydrophobic pharmaceutical compounds may be employed. Liposomes and emulsions are well known examples of delivery vehicles or carriers for hydrophobic drugs. Certain organic solvents such as dimethylsulfoxide also may be employed, although usually at the cost of greater toxicity. Additionally, the compounds may be delivered using a sustained-release system, such as semipermeable matrices of solid hydrophobic polymers containing the therapeutic agent. Various sustained-release materials have been established and are well known by those skilled in the art. Sustained-release capsules may, depending on their chemical nature, release the compounds for a few weeks up to over 100 days. Depending on the chemical nature and the biological stability of the therapeutic reagent, additional strategies for protein stabilization may be employed.

The pharmaceutical compositions also may comprise suitable solid or gel phase carriers or excipients. Examples of such carriers or excipients include but are not limited to calcium carbonate, calcium phosphate, various sugars, starches, cellulose derivatives, gelatin, and polymers such as polyethylene glycols.

Many of the protease modulating compounds of the invention may be provided as salts with pharmaceutically compatible counterions. Pharmaceutically compatible salts may be formed with many acids, including but not limited to hydrochloric, sulfuric, acetic, lactic, tartaric, malic, succinic, etc. Salts tend to be more soluble in aqueous or other protonic solvents that are the corresponding free base forms.

C. Effective Dosage

Pharmaceutical compositions suitable for use in the present invention include compositions where the active ingredients are contained in an amount effective to achieve its intended purpose. More specifically, a therapeutically effective amount means an amount of compound effective to prevent, alleviate or ameliorate symptoms of disease or prolong the survival of the subject being treated. Determination of a therapeutically effective amount is well within the capability of those skilled in the art, especially in light of the detailed disclosure provided herein.

For any compound used in the methods of the invention, the therapeutically effective dose can be estimated initially from cell culture assays. For example, a dose can be formulated in animal models to achieve a circulating concentration range that includes the IC_{50} as determined in cell culture (*i.e.*, the concentration of the test compound which achieves a half-maximal inhibition of the protease activity). Such information can be used to more accurately determine useful doses in humans.

Toxicity and therapeutic efficacy of the compounds described herein can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, *e.g.*, for determining the LD_{50} (the dose lethal to 50% of the population) and the ED_{50} (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio between LD_{50} and ED_{50}. Compounds which exhibit high therapeutic indices are preferred. The data obtained from these cell culture assays and animal studies can be used in formulating a range of dosage for use in human. The dosage of such compounds lies preferably within a range of circulating concentrations that include the ED50 with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. The exact formulation, route of administration and dosage can be chosen by the individual physician in view of the patient's condition. (See *e.g.*, Fingl *et al.*, 1975, in The Pharmacological Basis of Therapeutics, Ch. 1 p.1).

Dosage amount and interval may be adjusted individually to provide plasma levels of the active moiety which are sufficient to maintain the protease modulating effects, or minimal effective concentration (MEC). The MEC will vary for each compound but can be estimated from *in vitro* data; *e.g.*, the concentration necessary to achieve 50-90% inhibition of the protease using the assays described herein. Dosages necessary to achieve the MEC will depend on individual characteristics and route of administration. However, HPLC assays or bioassays can be used to determine plasma concentrations.

Dosage intervals can also be determined using MEC value. Compounds should be administered using a regimen which maintains plasma levels above the MEC for 10-90% of the time, preferably between 30-90% and most preferably between 50-90%.

In cases of local administration or selective uptake, the effective local concentration of the drug may not be related to plasma concentration.

The amount of composition administered will, of course, be dependent on the subject being treated, on the subject's weight, the severity of the affliction, the manner of administration and the judgment of the prescribing physician.

D. Packaging

The compositions may, if desired, be presented in a pack or dispenser device which may contain one or more unit dosage forms containing the active ingredient. The pack may for example comprise metal or plastic foil, such as a blister pack. The pack or dispenser device may be accompanied by instructions for administration. The pack or dispenser may also be accompanied with a notice associated with the container in form prescribed by a governmental agency regulating the manufacture, use, or sale of pharmaceuticals, which notice is reflective of approval by the agency of the form of the polynucleotide for human or veterinary administration. Such notice, for example, may be the labeling approved by the U.S. Food and Drug Administration for prescription drugs, or the approved product insert. Compositions comprising a compound of the invention formulated in a compatible pharmaceutical carrier may also be prepared, placed in an appropriate container, and labeled for treatment of an indicated condition. Suitable conditions indicated on the label may include treatment of a tumor, inhibition of angiogenesis, treatment of fibrosis, diabetes, and the like.

Functional Derivatives

Also provided herein are functional derivatives of a polypeptide or nucleic acid of the invention. By "functional derivative" is meant a "chemical derivative," "fragment," or "variant," of the polypeptide or nucleic acid of the invention, which

terms are defined below. A functional derivative retains at least a portion of the function of the protein, for example reactivity with an antibody specific for the protein, enzymatic activity or binding activity mediated through noncatalytic domains, which permits its utility in accordance with the present invention. It is well known in the art that due to the degeneracy of the genetic code numerous different nucleic acid sequences can code for the same amino acid sequence. Equally, it is also well known in the art that conservative changes in amino acid can be made to arrive at a protein or polypeptide that retains the functionality of the original. In both cases, all permutations are intended to be covered by this disclosure.

Included within the scope of this invention are the functional equivalents of the herein-described isolated nucleic acid molecules. The degeneracy of the genetic code permits substitution of certain codons by other codons that specify the same amino acid and hence would give rise to the same protein. The nucleic acid sequence can vary substantially since, with the exception of methionine and tryptophan, the known amino acids can be coded for by more than one codon. Thus, portions or all of the genes of the invention could be synthesized to give a nucleic acid sequence significantly different from one selected from the group consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35. The encoded amino acid sequence thereof would, however, be preserved.

In addition, the nucleic acid sequence may comprise a nucleotide sequence which results from the addition, deletion or substitution of at least one nucleotide to the 5'-end and/or the 3'-end of the nucleic acid formula selected from the group

consisting of those set forth in SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35, or a derivative thereof. Any nucleotide or polynucleotide may be used in this regard, provided that its addition, deletion or substitution does not alter the amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70 which is encoded by the nucleotide sequence. For example, the present invention is intended to include any nucleic acid sequence resulting from the addition of ATG as an initiation codon at the 5'-end of the inventive nucleic acid sequence or its derivative, or from the addition of TTA, TAG or TGA as a termination codon at the 3'-end of the inventive nucleotide sequence or its derivative. Moreover, the nucleic acid molecule of the present invention may, as necessary, have restriction endonuclease recognition sites added to its 5'-end and/or 3'-end.

Such functional alterations of a given nucleic acid sequence afford an opportunity to promote secretion and/or processing of heterologous proteins encoded by foreign nucleic acid sequences fused thereto. All variations of the nucleotide

sequence of the protease genes of the invention and fragments thereof permitted by the genetic code are, therefore, included in this invention.

Further, it is possible to delete codons or to substitute one or more codons with codons other than degenerate codons to produce a structurally modified polypeptide, but one which has substantially the same utility or activity as the polypeptide produced by the unmodified nucleic acid molecule. As recognized in the art, the two polypeptides are functionally equivalent, as are the two nucleic acid molecules that give rise to their production, even though the differences between the nucleic acid molecules are not related to the degeneracy of the genetic code.

A "chemical derivative" of the complex contains additional chemical moieties not normally a part of the protein. Covalent modifications of the protein or peptides are included within the scope of this invention. Such modifications may be introduced into the molecule by reacting targeted amino acid residues of the peptide with an organic derivatizing agent that is capable of reacting with selected side chains or terminal residues, as described below.

Cysteinyl residues most commonly are reacted with α-haloacetates (and corresponding amines), such as chloroacetic acid or chloroacetamide, to give carboxymethyl or carboxyamidomethyl derivatives. Cysteinyl residues also are derivatized by reaction with bromotrifluoroacetone, chloroacetyl phosphate, N-alkylmaleimides, 3-nitro-2-pyridyl disulfide, methyl 2-pyridyl disulfide, p-chloromercuribenzoate, 2-chloromercuri-4-nitrophenol, or chloro-7-nitrobenzo-2-oxa-1,3-diazole.

Histidyl residues are derivatized by reaction with diethylprocarbonate at pH 5.5-7.0 because this agent is relatively specific for the histidyl side chain. Para-bromophenacyl bromide also is useful; the reaction is preferably performed in 0.1 M sodium cacodylate at pH 6.0.

Lysinyl and amino terminal residues are reacted with succinic or other carboxylic acid anhydrides. Derivatization with these agents has the effect of reversing the charge of the lysinyl residues. Other suitable reagents for derivatizing primary amine containing residues include imidoesters such as methyl

picolinimidate; pyridoxal phosphate; pyridoxal; chloroborohydride; trinitrobenzenesulfonic acid; O-methylisourea; 2,4 pentanedione; and transaminase-catalyzed reaction with glyoxylate.

Arginyl residues are modified by reaction with one or several conventional reagents, among them phenylglyoxal, 2,3-butanedione, 1,2-cyclohexanedione, and ninhydrin. Derivatization of arginine residues requires that the reaction be performed in alkaline conditions because of the high pKa of the guanidine functional group. Furthermore, these reagents may react with the groups of lysine as well as the arginine α-amino group.

Tyrosyl residues are well-known targets of modification for introduction of spectral labels by reaction with aromatic diazonium compounds or tetranitromethane. Most commonly, N-acetylimidizol and tetranitromethane are used to form O-acetyl tyrosyl species and 3-nitro derivatives, respectively.

Carboxyl side groups (aspartyl or glutamyl) are selectively modified by reaction with carbodiimide (R'-N-C-N-R') such as 1-cyclohexyl-3-(2-morpholinyl(4-ethyl) carbodiimide or 1-ethyl-3-(4-azonia-4,4-dimethylpentyl) carbodiimide. Furthermore, aspartyl and glutamyl residues are converted to asparaginyl and glutaminyl residues by reaction with ammonium ions.

Glutaminyl and asparaginyl residues are frequently deamidated to the corresponding glutamyl and aspartyl residues. Alternatively, these residues are deamidated under mildly acidic conditions. Either form of these residues falls within the scope of this invention.

Derivatization with bifunctional agents is useful, for example, for cross-linking the component peptides of the protein to each other or to other proteins in a complex to a water-insoluble support matrix or to other macromolecular carriers. Commonly used cross-linking agents include, for example, 1,1-bis(diazoacetyl)-2-phenylethane, glutaraldehyde, N-hydroxysuccinimide esters, for example, esters with 4-azidosalicylic acid, homobifunctional imidoesters, including disuccinimidyl esters such as 3,3'-dithiobis(succinimidylpropionate), and bifunctional maleimides such as bis-N-maleimido-1,8-octane. Derivatizing agents such as methyl-3-[p-

azidophenyl) dithiolpropioimidate yield photoactivatable intermediates that are capable of forming crosslinks in the presence of light. Alternatively, reactive water-insoluble matrices such as cyanogen bromide-activated carbohydrates and the reactive substrates described in U.S. Patent Nos. 3,969,287; 3,691,016; 4,195,128; 4,247,642; 4,229,537; and 4,330,440 are employed for protein immobilization.

Other modifications include hydroxylation of proline and lysine, phosphorylation of hydroxyl groups of seryl or threonyl residues, methylation of the α-amino groups of lysine, arginine, and histidine side chains (Creighton, T.E., Proteins: Structure and Molecular Properties, W.H. Freeman & Co., San Francisco, pp. 79-86 (1983)), acetylation of the N-terminal amine, and, in some instances, amidation of the C-terminal carboxyl groups.

Such derivatized moieties may improve the stability, solubility, absorption, biological half life, and the like. The moieties may alternatively eliminate or attenuate any undesirable side effect of the protein complex and the like. Moieties capable of mediating such effects are disclosed, for example, in Remington's Pharmaceutical Sciences, 18th ed., Mack Publishing Co., Easton, PA (1990).

The term "fragment" is used to indicate a polypeptide derived from the amino acid sequence of the proteins, of the complexes having a length less than the full-length polypeptide from which it has been derived. Such a fragment may, for example, be produced by proteolytic cleavage of the full-length protein. Preferably, the fragment is obtained recombinantly by appropriately modifying the DNA sequence encoding the proteins to delete one or more amino acids at one or more sites of the C-terminus, N-terminus, and/or within the native sequence. Fragments of a protein are useful for screening for substances that act to modulate signal transduction, as described herein. It is understood that such fragments may retain one or more characterizing portions of the native complex. Examples of such retained characteristics include: catalytic activity; substrate specificity; interaction with other molecules in the intact cell; regulatory functions; or binding with an antibody specific for the native complex, or an epitope thereof.

Another functional derivative intended to be within the scope of the present invention is a "variant" polypeptide which either lacks one or more amino acids or contains additional or substituted amino acids relative to the native polypeptide. The variant may be derived from a naturally occurring complex component by appropriately modifying the protein DNA coding sequence to add, remove, and/or to modify codons for one or more amino acids at one or more sites of the C-terminus, N-terminus, and/or within the native sequence. It is understood that such variants having added, substituted and/or additional amino acids retain one or more characterizing portions of the native protein, as described above.

A functional derivative of a protein with deleted, inserted and/or substituted amino acid residues may be prepared using standard techniques well-known to those of ordinary skill in the art. For example, the modified components of the functional derivatives may be produced using site-directed mutagenesis techniques (as exemplified by Adelman *et al.*, 1983, *DNA* 2:183) wherein nucleotides in the DNA coding the sequence are modified such that a modified coding sequence is modified, and thereafter expressing this recombinant DNA in a prokaryotic or eukaryotic host cell, using techniques such as those described above. Alternatively, proteins with amino acid deletions, insertions and/or substitutions may be conveniently prepared by direct chemical synthesis, using methods well-known in the art. The functional derivatives of the proteins typically exhibit the same qualitative biological activity as the native proteins.

TABLES AND DESCRIPTION THEROF

This patent describes novel protease identified in databases of genomic sequence. The results are summarized in four tables, which are described below.

Table 1 documents the name of each gene, the classification of each gene, the positions of the open reading frames within the sequence, and the length of the corresponding peptide. From left to right the data presented is as follows: "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family",

"NA_length", "ORF Start", "ORF End", "ORF Length", and "AA_length". "Gene name" refers to name given the sequence encoding the protease enzyme. Each gene is represented by "SGPr" designation followed by an arbitrary number. The SGPr name usually represents multiple overlapping sequences built into a single contiguous sequence (a "contig"). The "ID#na" and "ID#aa" refer to the identification numbers given each nucleic acid and amino acid sequence in this patent application. "FL/Cat" refers to the length of the gene, with FL indicating full length, and "Cat' indicating that only the catalytic domain is presented. "Partial" in this column indicates that the sequence encodes a partial catalytic domain. "Superfamily" identifies whether the gene is a protease. "Group" and "Family" refer to the protease classification defined by sequence homology. "NA_length" refers to the length in nucleotides of the corresponding nucleic acid sequence. "ORF start" refers to the beginning nucleotide of the open reading frame. "ORF end" refers to the last nucleotide of the open reading frame, including the stop codon. "ORF length" refers to the length in nucleotides of the open reading frame (including the stop codon). "AA length" refers to the length in amino acids of the peptide encoded in the corresponding nucleic acid sequence.

Gene Name	ID#na	ID#aa	FL/Cat	Superfamily	Group	Family	NA_length	ORF Start	ORF End	ORF Length	AA_length
SGPr140	1	36	FL	Protease	Aspartyl	PepsinA1	1140	1	1140	1140	379
SGPr197	2	37	FL	Protease	Aspartyl	PepsinA1	1500	1	1500	1500	499
SGPr005	3	38	FL	Protease	Aspartyl	PepsinA1	1173	1	1173	1173	390
SGPr078	4	39	FL	Protease	Aspartyl	PepsinA1	1239	1	1239	1239	412
SGPr084	5	40	FL	Protease	Cysteine	HH	1191	1	1191	1191	396
SGPr009	6	41	FL	Protease	Cysteine	ICEp10	1137	1	1137	1137	378
SGPr286	7	42	Cat	Protease	Cysteine	ICEp20	705	1	705	705	234
SGPr008	8	43	FL	Protease	Cysteine	PepC2	2010	1	2010	2010	669
SGPr198	9	44	FL	Protease	Cysteine	PepC2	2112	1	2112	2112	703
SGPr210	10	45	FL	Protease	Cysteine	PepC2	2127	1	2127	2127	708
SGPr290	11	46	FL	Protease	Cysteine	PepC2	2136	1	2136	2136	711
SGPr116	12	47	FL	Protease	Cysteine	PepC2	2109	1	2109	2109	702
SGPr003	13	48	FL	Protease	Cysteine	PepC2	1542	1	1542	1542	513
SGPr016	14	49	partial	Protease	Metalloprotease	ADAM	846	1	846	846	281
SGPr352	15	50	FL	Protease	Metalloprotease	ADAM	3312	1	3312	3312	1103
SGPr050	16	51	FL	Protease	Metalloprotease	ADAM	3675	1	3675	3675	1224
SGPr282	17	52	FL	Protease	Metalloprotease	ADAM	2196	1	2196	2196	731
SGPr046	18	53	FL	Protease	Metalloprotease	ADAM	2805	1	2805	2805	934
SGPr060	19	54	FL	Protease	Metalloprotease	ADAM	4287	1	4287	4287	1428
SGPr068	20	55	FL	Protease	Metalloprotease	ADAM	3561	1	3561	3561	1186
SGPr096	21	56	FL	Protease	Metalloprotease	ADAM	5808	1	5808	5808	1935
SGPr119	22	57	FL	Protease	Metalloprotease	ADAM	4518	1	4518	4518	1505
SGPr143	23	58	FL	Protease	Metalloprotease	ADAM	2649	1	2649	2649	882
SGPr164	24	59	Cat	Protease	Metalloprotease	ADAM	2937	1	2937	2937	978
SGPr281	25	60	Cat	Protease	Metalloprotease	ADAM	3285	1	3285	3285	1094
SGPr075	26	61	Partial	Protease	Metalloprotease	ADAM	375	1	375	375	125
SGPr292	27	62	FL	Protease	Metalloprotease	PepM10	1710	1	1710	1710	569
SGPr069	28	63	FL	Protease	Metalloprotease	PepM13	2232	1	2232	2232	743
SGPr212	29	64	FL	Protease	Metalloprotease	PepM1	2730	1	2730	2730	909
SGPr049	30	65	FL	Protease	Metalloprotease	PepM1	2973	1	2973	2973	990
SGPr026	31	66	FL	Protease	Metalloprotease	PepM1	1953	1	1953	1953	650
SGPr203	32	67	FL	Protease	Metalloprotease	PepM1	2175	1	2175	2175	724
SGPr157	33	68	FL	Protease	Metalloprotease	PepM20	1524	1	1524	1524	507
SGPr154	34	69	FL	Protease	Metalloprotease	PepM20	1422	1	1422	1422	473
SGPr088	35	70	FL	Protease	Metalloprotease	PepM20	1428	1	1428	1428	475

Table 2 lists the following features of the genes described in this patent application: chromosomal localization, single nucleotide polymorphisms (SNPs), representation in dbEST, and repeat regions. From left to right the data presented is as follows: "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family", "Chromosome", "SNPs", "dbEST_hits", & "Repeats". The contents of the first 7 columns (i.e.,. "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family") are as described above for Table 1. "Chromosome" refers to the cytogenetic localization of the gene. Information in the "SNPs" column describes the nucleic acid position and degenerate nature of candidate single nucleotide polymorphisms (SNPs; please see table of polymorphism below). These SNPs were identified by blastn of the DNA sequence against the database of single nucleotide polymorphisms maintained at NCBI (http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html). "dbEST hits" lists accession numbers of entries in the public database of ESTs (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/index.html) that contain at least 150 bp of 100% identity to the corresponding gene. These ESTs were identified by blastn of dbEST. "Repeats" contains information about the location of short sequences, approximately 20 bp in length, that are of low complexity and that are present in several distinct genes.

Gene Name	ID#na	ID#aa	FL/Cat	Superfamily	Group	Family	Chromosome	SNPs	dbEST_hits	Repeats
SGPr140	1	36	FL	Protease	Aspartyl	PepsinA1	1p13-p33	ctggtggggcctggy, ss2008313_allelePos=201 ; ctctgtctactgcaacagk, ss703383_allelePos=201	A969042, AA411567	295 tgggtgccctctgtctactgc 315
SGPr197	2	37	FL	Protease	Aspartyl	PepsinA1	6p21.1	none	BF727344, BG394217, AW297327	None
SGPr005	3	38	FL	Protease	Aspartyl	PepsinA1	1p33	none	none	None
SGPr078	4	39	FL	Protease	Aspartyl	PepsinA1	11p15	aagtactcccaggy, ss20182_allelePos=101	BG260401, BF025894, BF793219	None
SGPr084	5	40	FL	Protease	Cysteine	HH	12q11	none	none	None
SGPr009	6	41	Cat	Protease	Cysteine	ICEp10	11q22	tgatggaaaataatgtr, ss726380_allelePos=201; gagacagctcaaay, ss866796_allelePos=187	none	900 cttcattgctttcaaatcttcc 921; 77 ttgatgatttgatggaaaat 96
SGPr286	7	42	FL	Protease	Cysteine	ICEp20	16p13.3	ytatgtggcctatcgcgatg; rs551848_allelePos=3135	none	574 ctgggagctgctgactgagg 592; 388 gtgggggcccacagctctcc 408
SGPr008	8	43	FL	Protease	Cysteine	PepC2	2p23	rccgaatgggagagggcn, ss678494_allelePos=201	BE075751	None
SGPr198	9	44	FL	Protease	Cysteine	PepC2	1q42.11	none	BE047777, AW339160	None
SGPr210	10	45	FL	Protease	Cysteine	PepC2	19q13.2	ggttccttgcaggy, ss1376193_allelePos=473	BE872274	1180 gaggaggatgacgaggatgagg 1201
SGPr290	11	46	FL	Protease	Cysteine	PepC2	2p23	none	none	1835 agcaggtgcacgctgccatg 1854
SGPr116	12	47	FL	Protease	Cysteine	PepC2	6p12	none	none	1003 ctggagatctgcaacctcac 1022
SGPr003	13	48	FL	Protease	Cysteine	PepC2	2q37	none	AL526645, BG475966, AL529373	1520 gctgctgcaggagccgctgctg 1541
SGPr016	14	49	partial	Protease	Metalloprotease	ADAM	8p11.1	none	AW569885, AI024863	710 ttaaatatatttcttctcataa 731
SGPr352	15	50	FL	Protease	Metalloprotease	ADAM	19p13.3	none	AW027573, AI131032, AI193804	1335 agactcgggccctgggctct 1354
SGPr050	16	51	FL	Protease	Metalloprotease	ADAM	5q15.3	tcggctgaaagggcy, ss1483925_allelePos=216	BF933693	2067 tttctctttctttgtcaa 2086; 2061 atttgatttctctttttctt 2080
SGPr282	17	52	FL	Protease	Metalloprotease	ADAM	16p12.3	ggcaatataaaagggcy, ss679422_allelePos=201; actcactgggctay, ss647742_allelePos=201; ggccggagcccaacgcaay, ss1226992_allelePos=101	none	None
SGPr046	18	53	FL	Protease	Metalloprotease	ADAM	18q23	none	none	2353 gtgaggaagagggagatgaagt 2374
SGPr060	19	54	FL	Protease	Metalloprotease	ADAM	15q26	none	AW575922, AW341169	None
SGPr068	20	55	FL	Protease	Metalloprotease	ADAM	10q22	none	AJ403134,	None
SGPr096	21	56	FL	Protease	Metalloprotease	ADAM	3p14	none	BE164543, AW995949, BF842288	None
SGPr119	22	57	FL	Protease	Metalloprotease	ADAM	12q11-q12	none	AU132053,	1257 tagagaaatgaaagttacaaa 1277
SGPr143	23	58	FL	Protease	Metalloprotease	ADAM	20p13	ggcagtggctactgcy, ss787708_allelePos=201	AA442551	2212 tgccactgtgctccaggctg 2231
SGPr164	24	59	Cat	Protease	Metalloprotease	ADAM	11q25	ataccgatcctgcaay, ss76755_allelePos=87	none	None
SGPr281	25	60	Cat	Protease	Metalloprotease	ADAM	5q31	none	none	None
SGPr075	26	61	Partial	Protease	Metalloprotease	ADAM	na	none	none	None
SGPr292	27	62	FL	Protease	Metalloprotease	PepM10	10q26	none	AW665196	52 gctcctggcccacccagcc 71; 959 aagcaattcaaaagctgtatg 979
SGPr089	28	63	FL	Protease	Metalloprotease	PepM13	Chr. 1	none	none	None
SGPr212	29	64	FL	Protease	Metalloprotease	PepM1	9q22	none	AL523882, T11456	None
SGPr049	30	65	FL	Protease	Metalloprotease	PepM1	5q21-q23	none	AI222989	2269 aatttaatatggaatatttat 2289
SGPr028	31	66	FL	Protease	Metalloprotease	PepM1	1q31-q36	none	none	None
SGPr203	32	67	FL	Protease	Metalloprotease	PepM1	2q37	none	AU132908, BE735172, BE563549 (many)	83 tggacgtggcctcggcctcca 103
SGPr157	33	68	FL	Protease	Metalloprotease	PepM20	18q22.3	none	BE386438, BE386547, BF920454 (many)	614 ccctggaggaacttgtggaa 633; 561 tcctgtgaatatcaaattca 580
SGPr154	34	69	FL	Protease	Metalloprotease	PepM20	1q32.1	gtcatctatggty, ss1269877_allelePos=223	none	806 tccttgcagctgctgtcagc 825
SGPr088	35	70	FL	Protease	Metalloprotease	PepM20	18q23	none	AL541127, AL542184, AL529661 (many)	None

Table 3 lists the extent and the boundaries of the protease catalytic domains, and other protein domains. The column headings are: "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Profile_start", "Profile_end", "Protease_start", "Protease_end", "Profile", and "Other Domains". The contents of the first 7 columns (i.e.,. "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family") are as described above for Table 1. "Profile Start", "Profile End", "Protease Start" and "Protease End" refer to data obtained using a Hidden-Markov Model to define catalytic range boundaries. The boundaries of the catalytic domain within the overall protein are noted in the "Protease Start" and "Protease End" columns. "Profile" indicates whether the HMMR search was done with a complete ("Global") or Smith Waterman ("Local") model, as described below. Starting from a multiple sequence alignment of catalytic domains, two hidden Markov models were built. One of them allows for partial matches to the catalytic domain; this is a "local" HMM, similar to Smith-Waterman alignments in sequence matching. The other model allows matches only to the complete catalytic domain; this is a "global" HMM similar to Needleman-Wunsch alignments in sequence matching. The Smith Waterman local model is more specific, allowing for fragmentary matches to the catalytic domain whereas the global "complete" model is more sensitive, allowing for remote homologue identification. The "Other domains" column lists the names and positions of domains within the protein sequence in addition to the protein protease domain. These domains were identified using PFAM (http://pfam.wustl.edu/hmmsearch.shtml) models, a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. Version 5.5 of Pfam (Sept 2000) contains alignments and models for 2478 protein families (http://pfam.wustl.edu/faq.shtml). The PFAM alignments were downloaded from http://pfam.wustl.edu/hmmsearch.shtml and the HMMr searches were run locally on a Timelogic computer (TimeLogic Corporation, Incline Village, NV).

Gene Name	ID#na	ID#aa	FL/Cat	Profile_start	Profile_end	Protease_start	Protease_end	Profile (global)	Other Domains
SGPr140	1	36	FL	1	356	65	378	Eukaryotic aspartyl protease	none
SGPr197	2	37	FL	1	32	199	230	Ubiquitin carboxyl-terminal hydrolases family 2	Zn-finger in ubiquitin-hydrolases (amino acid 26 to amino acid 96).
SGPr005	3	38	FL	1	358	65	389	Eukaryotic aspartyl protease	none
SGPr078	4	39	FL	1	356	70	409	Eukaryotic aspartyl protease	none
SGPr084	5	40	FL	1	163	23	185	Hedgehog amino-terminal signaling domain	Hint module amino acids 188-396.
SGPr009	6	41	Cat	1	141	131	264	ICE-like protease (caspase) p20 domain	ICE-like protease (caspase) p10 domain, amino acids 291-376; profile from 1-95. Also Caspase recruitment domain from amino acids 2-91.
SGPr286	7	42	FL	22	61	19	58	ICE-like protease (caspase) p20 domain	ICE-like protease (caspase) p10 domain, amino acids 144-202; profile from 1-61.
SGPr008	8	43	FL	2	344	35	333	Calpain family cysteine protease; Peptidase_C2	none
SGPr198	9	44	FL	1	344	45	344	Calpain family cysteine protease; Peptidase_C2	Calpain large subunit, domain III, amino acids 355-512, profile from 1-163. Also three EF hand motifs at amino acids 579-607, 609-637 and 674-701; all EF hands match from 1-26 of profile.
SGPr210	10	45	FL	1	344	45	341	Calpain family cysteine protease; Peptidase_C2	Calpain large subunit, domain III, amino acids 353-499, profile from 1-163. Also one EF hand motif at amino acids 613-641; EF hand matches from 1-26 of profile.
SGPr290	11	46	FL	1	344	43	346	Calpain family cysteine protease; Peptidase_C2	Calpain large subunit, domain III, amino acids 347-490, profile from 1-163. Also two EF hand motifs at amino acids 561-593 and 595-622; EF hands match from 1-26 of profile.
SGPr116	12	47	FL	1	344	42	341	Calpain family cysteine protease; Peptidase_C2	Calpain large subunit, domain III, amino acids 352-510, profile from 1-163. Also two EF hand motifs at amino acids 577-605 and 607-635; EF hands match from 1-26 of profile.
SGPr003	13	48	FL	1	344	13	322	Calpain family cysteine protease; Peptidase_C2	Calpain large subunit, domain III, amino acids 338-494, profile from 3-163.
SGPr016	14	49	partial	1	119	58	175	Reprolysin family propeptide, Pep_M12B_propep	none
SGPr352	15	50	FL	1	203	239	457	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, from amino acids 90-201, matching profile from 1-119. Also five Thrombospondin type 1 domains from 551-601, 829-884, 888-944, 946-1002, 1007-1057. All thrombospondin type 1 domains match profile from 1-54.
SGPr050	16	51	FL	3	203	292	495	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide from 111-235, matching profile from 1-119. Also has five Thrombospondin type 1 domains from 590-640, 930-986, 990-1047, 1055-1101, 1128-1180.
SGPr282	17	52	FL	1	119	75	190	Reprolysin family propeptide, Pep_M12B_propep	Disintegrin domain at amino acids 415-487; matche profile from 4-86. Also EGF-like domain at amino acids 633-661.
SGPr046	18	53	FL	13	203	1	194	Reprolysin (M12B) family zinc metalloprotease	Six Thrombospondin type 1 domains at 289-339, 569-627, 634-687, 689-736, 769-828, 844-890.
SGPr060	19	54	FL	1	203	639	860	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, Pep_M12B_propep from amino acids 502-615. Matches profile from 1-119. Also has one thrombospondin type 1 domain from 954-1004, matching profile from 1-54.
SGPr068	20	55	FL	3	203	261	460	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, Pep_M12B_propep from amino acids 120-240, matching profile from 1-119. Also has four thrombospondin type 1 domains between 556 - 1021.
SGPr096	21	56	FL	1	203	293	499	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, Pep_M12B_propep from amino acids 112-242, matching profile from 1-119. Also has 13 thrombospondin type 1 domains between 589 - 1733.
SGPr119	22	57	FL	1	203	259	467	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, Pep_M12B_propep from amino acids 92-215, matching profile from 1-119. Also has eight thrombospondin type 1 domains between 561 - 1418.
SGPr143	23	58	FL	1	203	275	478	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, Pep_M12B_propep from amino acids 145-263, matching profile from 1-119. Also has Disintegrin motif 495-570.
SGPr164	24	59	Cat	1	203	243	452	Reprolysin (M12B) family zinc metalloprotease	Reprolysin family propeptide, Pep_M12B_propep from amino acids 92-206, matching profile from 1-119. Also has three Thrombospondin type 1 domains from amino acids 545 to 978. Also has Glucose 6-phosphate dehydrogenase motif at 855-878.
SGPr281	25	60	Cat	89	203	317	432	Reprolysin (M12B) family zinc metalloprotease	Six Thrombospondin type 1 domains from amino acid 346 to 1030.
SGPr075	26	61	Partial	14	203	1	123	Reprolysin (M12B) family zinc metalloprotease	none
SGPr292	27	62	FL	1	171	56	267	Peptidase_M10, Matrixin	Also has four Hemopexin domains at amino acids 333-391, 394-449, 451-499, 508-549.
SGPr069	28	63	FL	1	225	535	742	Peptidase family M13	None
SGPr212	29	64	FL	343	374	275	306	Peptidase family M1	None
SGPr049	30	65	FL	1	441	98	506	Peptidase family M1	None
SGPr026	31	66	FL	1	441	32	417	Peptidase family M1	None
SGPr203	32	67	FL	161	441	194	444	Peptidase family M1	None
SGPr157	33	68	FL	42	368	106	450	Peptidase_M20	None
SGPr154	34	69	FL	1	247	55	288	Peptidase_M20	None
SGPr088	35	70	FL	1	368	22	417	Peptidase_M20	None

Table 4 describes the results of Smith Waterman similarity searches (Matrix: Pam100; gap open/extension penalties 12/2) of the amino acid sequences against the NCBI database of non-redundant protein sequences (http://www.ncbi.nlm.nih.gov/Entrez/protein.html). The column headings are: "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family", "Pscore", "aa_length", "aa_ID_match", "%Identity", "%Similar", "ACC#_nraa_match", and "Description". The contents of the first 7 columns (i.e.,. "Gene Name", "ID#na", "ID#aa", "FL/Cat", "Superfamily", "Group", "Family") are as described above for Table 1. "Pscore" refers to the Smith Waterman probability score. This number approximates the chance that the alignment occurred by chance. Thus, a very low number, such as 2.10E-64, indicates that there is a very significant match between the query and the database target. "aa_length" refers to the length of the protein in amino acids. "aa_ID_match" indicates the number of amino acids that were identical in the alignment. "% Identity" lists the percent of amino acids that were identical over the aligned region. "% Similarity" lists the percent of amino acids that were similar over the alignment. "ACC#nraa_match" lists the accession number of the most similar protein in the NCBI database of non-redundant proteins. "Description" contains the name of the most similar protein in the NCBI database of non-redundant proteins.

Smith Waterman

Gene Name	ID#n	ID#aa	FL/Cat	Superfamily	Group	Family	Pscore	aa_length	aa_ID_match	%Identity	%Similar	ACC#_nraa_match	Description
SGPr140	1	36	FL	Protease	Aspartyl	PepsinA1	1.4E-160	379	263	66	76	CAC19554.1	Chymosin [Camelus dromedarius]
SGPr197	2	37	FL	Protease	Aspartyl	PepsinA1	6.9E-137	499	296	46	56	CAB66759.1	Hypothetical histone deacetylase [Homo sapiens]
SGPr005	3	38	FL	Protease	Aspartyl	PepsinA1	1.4E-130	390	230	62	76	BAB11755.1	Pepsinogen C [Rhinolophus ferrumequinum]
SGPr078	4	39	FL	Protease	Aspartyl	PepsinA1	3.2E-285	412	412	100	100	NP_001900.1	Cathepsin D (lysosomal aspartyl protease) [Homo sapiens]
SGPr084	5	40	FL	Protease	Cysteine	HH	3E-259	396	396	100	100	O43323	DESERT HEDGEHOG PROTEIN PRECURSOR (DHH) (HHG-3) [Homo sapiens]
SGPr009	6	41	FL	Protease	Cysteine	ICEp10	3.5E-129	378	223	55	67	NP_033938.1	Caspase 12 [Mus musculus]
SGPr286	7	42	Cat	Protease	Cysteine	ICEp20	4.6E-42	234	108	46	65	NP_036246.1	Caspase 14, apoptosis-related cysteine protease [Homo sapiens]
SGPr008	8	43	FL	Protease	Cysteine	PepC2	9.1E-86	669	229	33	53	AAD34601.1	Lens-specific calpain Lp82 [Oryctolagus cuniculus]
SGPr198	9	44	FL	Protease	Cysteine	PepC2	0	703	593	84	92	BAA03369.1	Calpain [Rattus norvegicus]
SGPr210	10	45	FL	Protease	Cysteine	PepC2	0	708	569	79	86	CAC10066.1	Calpain 12 [Mus musculus]
SGPr290	11	46	FL	Protease	Cysteine	PepC2	6.2E-103	711	256	39	56	AAD34601.1	Lens-specific calpain Lp82 [Oryctolagus cuniculus]
SGPr116	12	47	FL	Protease	Cysteine	PepC2	0	702	702	100	100	NP_008989.1	Calpain 11 [Homo sapiens]
SGPr003	13	48	FL	Protease	Cysteine	PepC2	0	513	513	100	100	NP_075574.1	Calpain 10 [Homo sapiens]
SGPr016	14	49	partial	Protease	Metalloprotease	ADAM	1.3E-89	282	215	52	58	S47656	tMDC II (ADAM 5-like) protein - crab-eating macaque
SGPr352	15	50	FL	Protease	Metalloprotease	ADAM	0	1103	1072	100	100	AAG35563.1	Zinc metalloendopeptidase [Homo sapiens]
SGPr050	16	51	FL	Protease	Metalloprotease	ADAM	6.8E-149	1224	385	37	53	AAG35563.1	Zinc metalloendopeptidase [Homo sapiens]
SGPr282	17	52	FL	Protease	Metalloprotease	ADAM	0	731	619	85	91	I52361	Metalloproteinase-like, disintegrin-like, cysteine-rich protein IVa [crab-eating macaque]
SGPr046	18	53	FL	Protease	Metalloprotease	ADAM	1.1E-162	934	320	39	56	AAG35563.1	Zinc metalloendopeptidase [Homo sapiens]
SGPr060	19	54	FL	Protease	Metalloprotease	ADAM	5.2E-87	1428	250	39	55	NP_055087.1	Disintegrin-like and metalloprotease with thrombospondin type 1 motif, 7 [Homo sapiens]
SGPr068	20	55	FL	Protease	Metalloprotease	ADAM	0	1188	624	64	77	O15072	ADAM-TS 3 PRECURSOR [Homo sapiens]
SGPr098	21	56	FL	Protease	Metalloprotease	ADAM	0	1935	1465	100	100	BAA92550.1	KIAA1312 protein [Homo sapiens]
SGPr119	22	57	FL	Protease	Metalloprotease	ADAM	0	1505	699	53	70	BAA92550.1	KIAA1312 (ADAMS 9-like) protein [Homo sapiens]
SGPr143	23	58	FL	Protease	Metalloprotease	ADAM	0	882	726	99	99	CAC16509.2	Novel disintegrin and reprolysin metalloproteinase [Homo sapiens]
SGPr164	24	59	Cat	Protease	Metalloprotease	ADAM	1.8E-264	978	465	50	67	XP_012976.1	ADAMS-1 preproprotein [Homo sapiens]
SGPr281	25	60	Cat	Protease	Metalloprotease	ADAM	.4e-07	1094	287	39	55	NP_112217.1	ADAMTS 12 [Homo sapiens]
SGPr075	26	61	Partial	Protease	Metalloprotease	ADAM	1.1E-54	125	98	65	73	CAC18729	Metalloprotease/disintegrin [Rattus norvegicus]
SGPr292	27	62	FL	Protease	Metalloprotease	PepM10	6E-137	569	333	57	74	AAC21447.1	Matrix metalloproteinase [Xenopus laevis]
SGPr069	28	63	FL	Protease	Metalloprotease	PepM13	0	743	581	78	90	AAG18446.1	Neprilysin-like peptidase alpha [Mus musculus]
SGPr212	29	64	FL	Protease	Metalloprotease	PepM1	1.4E-31	909	55	77	87	BAB25647.1	Probable zinc metall proteinase [Mus musculus]
SGPr049	30	65	FL	Protease	Metalloprotease	PepM1	4.1E-220	990	375	68	79	BAB29490.1	Putative aminopeptidase [Mus musculus]
SGPr026	31	66	FL	Protease	Metalloprotease	PepM1	0	650	650	100	100	AAH01064	Hypothetical protein DKFZp547H084 [Homo sapiens]
SGPr203	32	67	FL	Protease	Metalloprotease	PepM1	1.9E-276	724	493	100	100	AAG22080.1	RNPEP-like protein [Homo sapiens]
SGPr157	33	68	FL	Protease	Metalloprotease	PepM20	7.5E-202	507	310	100	100	AAH04271.1	Hypothetical protein [Homo sapiens]
SGPr154	34	69	FL	Protease	Metalloprotease	PepM20	1.9E-28	473	122	31	48	AAK22721.1	M20/M25/M40 family peptidase [Caulobacter crescentus]
SGPr088	35	70	FL	Protease	Metalloprotease	PepM20	9.8e-315	475	475	100	100	XP_008819.1	Hypothetical protein FLJ10830 [Homo sapiens]

EXAMPLES

The examples below are not limiting and are merely representative of various aspects and features of the present invention. The examples below demonstrate the isolation and characterization of the proteases of the invention.

EXAMPLE 1: Identification of Genomic Fragments Encoding Proteases

Novel proteases were identified from the Celera human genomic sequence databases, and from the public Human Genome Sequencing project (http://www.ncbi.nlm.nih.gov/) using hidden Markov models (HMMR). The genomic database entries were translated in six open reading frames and searched against the model using a Timelogic Decypher box with a Field programmable array (FPGA) accelerated version of HMMR2.1. The DNA sequences encoding the predicted protein sequences aligning to the HMMR profile were extracted from the original genomic database. The nucleic acid sequences were then clustered using the Pangea Clustering tool to eliminate repetitive entries. The putative protease sequences were then sequentially run through a series of queries and filters to identify novel protease sequences. Specifically, the HMMR identified sequences were searched using BLASTN and BLASTX against a nucleotide and amino acid repository containing known human proteases and all subsequent new protease sequences as they are identified. The output was parsed into a spreadsheet to facilitate elimination of known genes by manual inspection. Two models were used, a "complete" model and a "partial" or Smith Waterman model. The partial model was used to identify sub-catalytic domains, whereas the complete model was used to identify complete catalytic domains. The selected hits were then queried using BLASTN against the public NRNA and EST databases to confirm they are indeed unique.

Extension of partial DNA sequences to encompass the longer sequences, including full-length open-reading frame, was carried out by several methods.

Iterative blastn searching of the cDNA databases listed in Table 5 was used to find cDNAs that extended the genomic sequences. "LifeGold" databases are from Incyte Genomics, Inc (http://www.incyte.com/). NCBI databases are from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). All blastn searches were conducted using a blosum62 matrix, a penalty for a nucleotide mismatch of –3 and reward for a nucleotide match of 1. The gapped blast algorithm is described in: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402).

Extension of partial DNA sequences to encompass the full-length open-reading frame was also carried out by iterative searches of genomic databases. The first method made use of the Smith-Waterman algorithm to carry out protein-protein searches of the closest homologue or orthologue to the partial. The target databases consisted of Genscan [Chris Burge and Sam Karlin "Prediction of Complete Gene Structures in Human Genomic DNA", JMB (1997) 268(1):78-94)] and open-reading frame (ORF) predictions of all human genomic sequence derived from the human genome project (HGP) as well as from Celera. The complete set of genomic databases searched is shown in Table 6 below. Genomic sequences encoding potential extensions were further assessed by blastp analysis against the NCBI nonredundant to confirm the novelty of the hit. The extending genomic sequences were incorporated into the cDNA sequence after removal of potential introns using the Seqman program from DNAStar. The default parameters used for Smith-Waterman searches were as shown next. Matrix: PAM100; gap-opening penalty: 12; gap extension penalty: 2. Genscan predictions were made using the Genscan program as detailed in Chris Burge and Sam Karlin "Prediction of Complete Gene Structures in Human Genomic DNA", JMB (1997) 268(1):78-94). ORF predictions from genomic DNA were made using a standard 6-frame translation.

Another method for defining DNA extensions from genomic sequence used iterative searches of genomic databases through the Genscan program to predict

exon splicing [Burge and Karlin, JMB (1997) 268(1):78-94)]. These predicted genes were then assessed to see if they represented "real" extensions of the partial genes based on homology to related proteases.

Another method involved using the Genewise program (http://www.sanger.ac.uk/Software/Wise2/) to predict potential ORFs based on homology to the closest orthologue/homologue. Genewise requires two inputs, the homologous protein, and genomic DNA containing the gene of interest. The genomic DNA was identified by blastn searches of Celera and Human Genome Project databases. The orthologs were identified by blastp searches of the NCBI non-redundant protein database (NRAA). Genewise compares the protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.

Table 5. Databases used for cDNA-based sequence extensions

Database	Database Date
LifeGold templates	March 2001
LifeGold compseqs	March 2001
LifeGold compseqs	March 2001
LifeGold compseqs	March 2001
LifeGold fl	March 2001
LifeGold flft	March 2001
NCBI human Ests	March 2001
NCBI murine Ests	March 2001
NCBI nonredundant	March 2001

I. TABLE 6. DATABASES USED FOR GENOMIC-BASED SEQUENCE EXTENSIONS

Database	Number of entries	Database Date
Celera v. 1-5	5,306,158	Jan 2000
Celera v. 6-10	4,209,980	March 2000
Celera v. 11-14	7,222,425	April 2000
Celera v. 15	243,044	April 2000
Celera v. 16-17	25,885	April 2000
Celera Assembly 5 (release 25h)	479,986	March 2001
HGP Phase 0	3,189	Nov 1/00
HGP Phase 1	20,447	Jan 1/01
HGP Phase 2	1,619	Jan 1/01
HGP Phase 3	9,224	March 2001
HGP Chromosomal assemblies	2759	March 2001

Results:

The sources for the sequence information used to identify genes are listed below. For genes that were extended using Genewise, the accession numbers of the protein ortholog and the genomic DNA are given. (Genewise uses the ortholog to assemble the coding sequence of the target gene from the genomic sequence). The amino acid sequences for the orthologs were obtained from the NCBI non-redundant database of proteins .(http://www.ncbi.nlm.nih.gov/Entrez/protein.html). The genomic DNA came from two sources: Celera and NCBI-NRNA, as indicated below. cDNA sources are also listed below. All of the genomic sequences were used as input for Genscan predictions to predict splice sites [Burge and Karlin, JMB

(1997) 268(1):78-94)]. Abbreviations: HGP: Human Genome Project; NCBI, National Center for Biotechnology Information.

SGPr140, SEQ ID NOS:1, 36
Genomic DNA source: Celera Assembly 5h contig 90000642234645
Homologs used for Genewise: gi_5822085, gi_11265696, gi_2136604

SGPr197, SEQ ID NOS:2, 37
Genomic DNA source: Celera Assembly 5h contig 90000640151915
Homologs used for Genewise: gi_12731929, gb_AAA60062.1, gi_999902

SGPr005, SEQ ID NOS:3, 38
Genomic DNA source: Celera Assembly 5h contig 90000642234645
Homologs used for Genewise: gi_11265695, gi_12731929, dbj_BAB11754.1

The genomic sequence containing the original HMM hit was blast against Celera_Asm5h where it aligned with contig 90000642234645 (4157978 bp) in the anti-sense orientation. 200 kb of the contig was used for genewise/genscan/ sym4 predictions. Genewise was run with human pepsinogen C (gi|12731929) as the model and the result extended the original HMM hit to 370 aa. The genewise prediction was then blastx against NCBI_nonredundant to find that it shared strongest homology (64% identity over 372 aa) with pepsinogen C from *Rhinolophus ferrumequinumxx*. The extended sequence also shares homology (74% over 324 aa) with the profiled Pfam Eukaryotic aspartyl protease. All overlapping Genscan predictions were blastx vs NCBI_nonredundant. Only one prediction (id 83280) contained sequence with homology to pepsinogen C. The genewise prediction was then blastn vs all EST and cDNA databases. Several hits were found:

1.) LGTemplatesMAR2001: AAA41827.1 g206083 pepsinogen 0
2.) LGcompseqsMAR2001: 7477287CB1
3.) LGcompseqsMAR2001: 825016H1
4.) LGflMAR2001: 7477287CB1

5.) LGflMAR2001n: g8546678_edit_02

6.) LGflMAR2001n: 825016H1_edit_1

7.) LGfltAPR2001n: 7477287CB1

The LGcompseqsMAR2001 EST 7477287CB1 contains an ORF of 1173 bp or 390 aa. When blastx against NCBI_nonredundant 7477287CB1 shares 62% identity over 372 aa to pepsinogen C of *Rhinolophus ferrumequinum*. When aligned with the SGPr005 genewise prediction, 7477287CB1 has 3 conflicts and 4 inserts/deletions.

Conflict #1

The first conflict occurs at nucleotide 189 of 7477287CB1. In the 7477287CB1 sequence nucleotide 189 is a "T" while in the SGPr005 genewise prediction the corresponding nucleotide is a "C". The nucleotide conflict is silent and does not give rise to an amino acid change.

At conflict #1 both the SGPr005 genewise sequence and the 7477287CB1 sequence are supported by genomic data.

SGPr005 genewise sequence: Celera_Asm5h contig 90000642234645

7477287CB1: HGP_s contig gi|9213869_5

Conflict #2

The second conflict occurs at nucleotide 379 of 7477287CB1. In the 7477287CB1 sequence nucleotide 379 is a "G" while in the SGPr005 genewise prediction the corresponding nucleotide is a "A". The nucleotide conflict gives rise to an aa change of D (7477287CB1) to N (SGPr005).

At conflict #2 both the SGPr005 genewise sequence and the 7477287CB1 sequence are supported by genomic data.

SGPr005 genewise sequence: Celera_Asm5h contig 90000642234645

7477287CB1: HGP_s contig gi|9213869_5

Conflict #3

The third conflict occurs at nucleotide 745 of 7477287CB1. In the 7477287CB1 sequence nucleotide 745 is a "G" while in the SGPr005 genewise

prediction the corresponding nucleotide is a "T". The nucleotide conflict gives rise to an aa acid conflict of E (7477287CB1) to STOP (SGPr005).

At conflict #3 the only sequence supported by genomic data is the SGPr005 genewise sequence which gives rise to the stop codon.

SGPr005 genewise sequence: Celera_Asm5h contig 90000642234645 and HGP_s contig gi|9213869_5

Inserts #1 and #2

The first two inserts occurs at nucleotide 214 of the SGPr005 genewise predicted sequence and nucleotide 297 of 7477287CB1.

7477287CB1: TTCCTAGTC
_TCTTTGATACGGGTTCCTCCAATCTGTAGCCTGCCCTC

SGPr005gw:
TTCCTAGTCCTCTTTGATACGGGTTCCTCCAATCTGTAG_ CTGCCCTC

Because one insert occurs on the genewise prediction while the other occurs on the EST the two sequences are only frameshifted for 31 nucleotides. When this stretch of sequence is blastx vs NCBI_nonredundant, it is clear that the SGPr005 genewise predicted sequence contains the correct reading frame in order to maintain homology to pepsinogen C.

The genomic data from Celera_Asm5h contig 90000642234645 supports the SGPr005 genewise sequence while the HGP_s contig gi|9213869_5 supports the 7477287CB1 sequence.

Insert #3

The third insert occurs at nucleotide 706 of 7477287CB1.

7477287CB1:
ATCCTTGGAGGTGTGGACCCCAACCTTTATTCTGGTCAGATCATCTGGACC

SGPr005gw: ATCCTTGGAGGTGTGGACCCCAAC_
TTTATTCTGGTCAGATCATCTGGACC

When this stretch of sequence is translated and blastp vs ncbi_redundant, it is clear that the 7477287CB1 sequence contains the necessary reading frame to maintain homology with pepsinogen C. However, both the Celera_Asm5h and

HGP_s genomic hits (Celera_Asm5h contig 90000642234645 and HGP_s contig gi|9213869_5) support the SGPr005 genewise predicted sequence.

Insert #4

The fourth insert occurs at nucleotide 873 of 7477287CB1.

7477287CB1:

GAGACCTTCCTGCTGGCAGTTCCTCAGCAGTACATGGCCTCCTTCCTGCAG

SGPr005gw: GAGACCTTCCTGCTGGCAGTTCCTCAGCAGTACAT_GCCTCCTTCCTGCAG

When this stretch of sequence is translated and blastp vs ncbi_redundant, it is clear that the 7477287CB1 sequence contains the necessary reading frame to maintain homology with pepsinogen C. However, both the Celera_Asm5h and HGP_s genomic hits (Celera_Asm5h contig 90000642234645 and HGP_s contig gi|9213869_5) support the SGPr005 genewise predicted sequence.

SGPr078, SEQ ID NOS:4, 39
Genomic DNA source: Public genomic contig: gi|11560222, subfragment 11
Homologs used for Genewise: gi_5822085

SGPr084, SEQ ID NOS:5, 40
Genomic DNA source: Celera Assembly 5h contig 90000636191372
Homologs used for Genewise: gb_AAD31927.1, sp_O43323, ref_NP_031883.1

SGPr009, SEQ ID NOS:6, 41
Genomic DNA source: Celera Assembly 5h contig 90000642045264
Homologs used for Genewise: gi_12736472, gb_AAC99852.1, gb_AAC99854.1

The original HMM hit was blast against Celera_Asm5h where it aligned with contig 90000642045264 (8,329,407bp) in the sense orientation. Nucleotides 14,659 to 111,952 of the contig were used for genewise/genscan/sym4 predictions. Genewise was run with human caspase 4 (gi|12736472|gn1) as the model and the

prediction extended SGPr009 through the 3' most 274aa (through the stop codon). The SGPr009 genewise prediction shares homology (62% identity over 274 aa) with human caspase 4 (gi|4502577). The genewise prediction also overlaps SGPr111, merging these two fragments into one gene (SGPr009=SGPr111). However, the genewise prediction does have one internal stop and one frame shift. The internal stop codon and the frame shift were corrected for through analysis with other genomic contigs and ESTs. One EST of importance was LGcompseqsMAR2001 7478251CB1 which overlaps with the SGPr009 genewise prediction and extends the prediction in the 5' direction through the start codon. To correct for sequencing errors in the extended 7478251CB1 sequence, the EST was blastn vs. genomic databases and the following changes were made: nucleotide 391 and 393 were changed from A to G based on HGP_s and Celera contigs, and nucleotide 1041 was changed from A to T based on HGP_s and Celera contigs.

SGPr286, SEQ ID NOS:7, 42

Genomic DNA source: Celera Assembly 5h contig 90000628729589

Homologs used for Genewise: ref_NP_036246.1, gi_6753280

The genomic sequence containing the original HMM hit was blast against Celera_Asm5h where it aligned with contig 90000628729589 (1,488,284 bp) in the anti-sense orientation. 200 kb of the contig was used for genewise/genscan/sym4 predictions. Genewise was run with human caspase 14 (gi|6912286) as the model and the result extended the original HMM hit to 233 aa. The genewise result shares good homology to caspase 14 (44% identity over 236aa) from amino acid 11 through the stop codon. The genewise result was then blastn vs. all EST and cDNA databases where it hit several ESTs: LGtemplatesMAR2001:

292606.4, LGfliftAPR2001n: 7648238CB1, LGcompseqsMAR2001: 7648638J1, 7013516H1, NCBI Nonredundant NA: gi|3982609, mega_cdna: cluster381375_2_incyte, cluster381375_-4_incyte. The overlapping EST data was used to support the genewise prediction.

SGPr008, SEQ ID NOS:8, 43
Genomic DNA source: Celera Assembly 5h contig 301714258
Homologs used for Genewise: emb_CAA86994.1, gb_AAF57563.1, gb_AAF57564.1

SGPr198, SEQ ID NOS:9, 44
Genomic DNA source: Celera Assembly 5h contigs: 9802310, 90000642810957
Homologs used for Genewise: gb_AAF99682.1, gb_AAG22771.1, gi_12722673

SGPr210, SEQ ID NOS:10, 45
Genomic DNA source: Celera Assembly 5h contig 92000004252572
Homologs used for Genewise: emb_CAC10067.1, emb_CAC10068.1, ref_NP_068694.1

SGPr290, SEQ ID NOS:11, 46
Genomic DNA source: Celera Assembly 5h contig 301714258
Homologs used for Genewise: gb_AAD34600.1, gb_AAD51699.1, gb_AAD56236.1

SGPr116, SEQ ID NOS:12, 47
Genomic DNA source: Celera Assembly 5h contig 90000627067487
Homologs used for Genewise: sp_P00789, gi_12732105, ref_NP_008989.1

SGPr003, SEQ ID NOS:13, 48
Genomic DNA source: 90000640081635
Homologs used for Genewise: gb_AAH05681.1, ref_NP_035926.1, gb_AAG17967.1
Notes: Recently published as ref|NP_075574.1| calpain 10, isoform d; calcium-activated neutral protease

SGPr016, SEQ ID NOS:14, 49
Genomic DNA source: Celera Assembly 5h contig 90000642821147
Homologs used for Genewise: gi_1079470, ref_NP_055052.1

Notes: Genomic region may be misassembled, predicted protein may have gaps in the middle. Used incyte template 094916.1 to extend genewise prediction

SGPr352, SEQ ID NOS:15, 50
Genomic DNA source: Celera Assembly 5h contig 90000628457498
Homologs used for Genewise: ref_NP_055087.1, gb_AAG35563.1, gb_AF163762.1

SGPr050, SEQ ID NOS:16, 51
Genomic DNA source: Celera Assembly 5h contig 90000626814267
Homologs used for Genewise: ref_NP_055087.1, gb_AAG35563.1
Used Incyte sequences to aid gene finding and show tissue expression: 333039.1, 333039.4, 1011933.1, 333039.3, 333039.2, 3533147CB1. Clones were expressed in urinary tract (9), respiratory system (3), female genitalia (2), nervous system (2) and connective, exocrine, digestive and musculoskeletal systems (one each)

SGPr282, SEQ ID NOS:17, 52
Genomic DNA source: Celera Assembly 5h contig 90000641115460
Homologs used for Genewise: gb_AAC09475.1, pir_|I65253

SGPr046, SEQ ID NOS:18, 53
Genomic DNA source: Celera Assembly 5h contig 92000004436076
Homologs used for Genewise: ref_NP_055087.1, gb_AAG35563.1
Also used Incyte sequences 207915.2 , 207915.5, 207915.11 , 207915.4, 7478405CB1, 9123702. Resolved differences between genomic and EST sequence by blasting against Celera raw reads, public and Incyte ESTs and HGP genomic contigs.

SGPr060, SEQ ID NOS:19, 54
Genomic DNA source: Celera Assembly 5h contig 90000642001297

Homologs used for Genewise: gb_AAG35563.1, ref_NM_022122.1, ref_NP_112217.1
Incyte sequences 452273.1, 013006.4, 013006.3, 322264.1 and public ESTs gi|7115818, gi|6837795 were used to extend and verify the genewise prediction.

SGPr068, SEQ ID NOS:20, 55
Genomic DNA source: Celera Assembly 5h contig 90000624770881
Homologs used for Genewise: sp_O15072, gi_11417111, gi_12731510
Incyte sequence 7477386CB1, 1719204CB1 also used. Sequence from 3062-3172 in the mRNA is missing in incyte sequence 7477386CB1, leading to the replacement of the peptide "GNHQNSTVRADVWELGTPEGQWVPQSEPLHPINKISST" with "A" in the predicted protein. In 7477386CB1 there are two 3nt inserts at splice sites, and a 5 nt insert followed shortly by a 1 nt insert, none of which are found in any genomic sequences, and so may be the result of atypical splicing. This alternative form would insert a V at position 291 of the protein, a Q at 318, a G at 386, and changes a LWS at 584-586 to a PAYGG. Incyte template 196583.5 uses an alternative splice acceptor site in one intron, inserting the sequence "CTCCCCATCTCCCCTCAG" at position 2420 of the mRNA and inserting the sequence PISPQA into the protein.

SGPr096, SEQ ID NOS:21, 56
Genomic DNA source: Celera Assembly 5h contig 90000637859600
Homologs used for Genewise: dbj_BAA92550.1, ref_NP_064634.1
Partial fragments published in 2000 as NP_064634.1 and as KIAA1312. 121 ESTs from Incyte template 1501550.6, show broad expression, highest in female genitalia and nervous system.

SGPr119, SEQ ID NOS:22, 57
Genomic DNA source: Celera Assembly 5h contig 90000642194924
Homologs used for Genewise: dbj_BAA92550.1, ref_NP_064634.1

Public sequence gi|13376516|ref|NM_025003.1 encodes an alternative splice form which is missing 3586–3693 of the RNA sequence

SGPr143, SEQ ID NOS:23, 58
Genomic DNA source: Celera Assembly 5h contig 90000641832427
Homologs used for Genewise: em_|CAC16509.2, gb_AAB51194.1, gb_AAK07852.1

SGPr164, SEQ ID NOS:24, 59
Genomic DNA source: Celera Assembly 5h contig 90000642493829
Homologs used for Genewise: sp_P97857, ref_NP_077376.1, dbj_BAA11088.1
3 ESTs cover this gene. 2 are from brain tumors, 1 from testis. One EST has 1 AA deletion. Start is probably at first Met in the AA sequence

SGPr281, SEQ ID NOS:25, 60
Genomic DNA source: Celera Assembly 5h contig 92000004763172
Homologs used for Genewise: emb_AL523577.1

SGPr075, SEQ ID NOS:26, 61
Genomic DNA source: Assembly of Celera Assembly 5g contigs 165000100324361, 165000102322372, 165000101460952, 165000102528372, 165000102358388, 165000100557102 , 165000102544200, 165000102496419 , 165000101581219 , 165000100483148, 165000100004880, 165000102322372, 165000100324361
Homologs used for Genewise: emb_CAC18729.1
The nnn in NA sequence and X in peptide sequence represents a probable missing exon; the gene may also be incomplete at either end. Based on searches of all human DNA databases, this gene is likely to be a fragment of the ortholog of the rat gene used as genewise homolog.

SGPr292, SEQ ID NOS:27, 62
Genomic DNA source: Celera Assembly 5h contig 90000641768196
Homologs used for Genewise: gb_AAH02631.1, ref_NP_077278.1, gb_AAC21447.1
The following polymorphisms are seen: C->T at 572, T->A at 591, T->A at 593, C->T at 981, deletion of A at 1720. The first is seen in some ESTs and public genomic sources and changes an A to a V in the protein; the second and third are seen in ESTs and a single public genomic sequence, and change a V to an E in the protein. The third is seen only in ESTs and is a synonymous substitution. The fourth is seen in ESTs and public genomic data and is in the 3' UTR of the gene. In addition, a 3nt deletion at 701-703 is seen in Incyte template 1510368.1, resulting in deletion of the D at position 558 of the peptide.

SGPr069, SEQ ID NOS:28, 63
Genomic DNA source: Celera Assembly 5h contig 90000624872437
Homologs used for Genewise: gb_AAG18446.1, gb_AAG18448.1, gb_AAF69247.1

SGPr212, SEQ ID NOS:29, 64
Genomic DNA source: Celera Assembly 5h contig 90000640657088
Homologs used for Genewise: dbj_BAB25647.1, pir_A75464, sp_P91885

SGPr049, SEQ ID NOS:30, 65
Genomic DNA source: Celera Assembly 5h contig 90000641091876
Homologs used for Genewise: dbj_BAB29490.1, emb_AL543134.1, sp_P15145, gb_AAC32807.1
An alternatively spliced form is predicted by public EST gi|3805192 in which an extra exon ("TCTTTTATTTACTTTTTTAACTACAGCCACACTTTGAGCAG") is inserted at position 3335 of the mRNA. This has an in-frame stop codon at it's end

and so predicts a truncated protein, which has the first 918 AA of the predicted protein, followed by "SLLFTFLTTATL*". Also used Incyte sequences 231695.1, 231695.7, 231695.2 to aid the prediction

SGPr026, SEQ ID NO:31, SEQ ID NO:66
Genomic DNA source: Celera Assembly 5h contig 113000081526387 and public genomic contig gi|12227482
Homologs used for Genewise: gi_12654473, gi_10933784, gi_10800858 (all parital seqs of this gene), gi_1754515 (rat ortholog)
gi|9368836 encodes an alternative splice form, missing one exon, and with another exon extended. It predicts a truncated protein product, with AA 1-230 of the main form, followed by EPGVG*.

SGPr203, SEQ ID NO:32, SEQ ID NO:67
Genomic DNA source: Celera Assembly 5h contig 90000640081635
Homologs used for Genewise: ref_NM_016552.1, emb_CAC14047.1, gb_AAG22080.1
A splice variant is created by use of an alternative splice acceptor that eliminates from 1526-1558 in ESTs such as Incyte cDNA 1868183CA2, resulting in the removal of the peptide LEFERWLNATG from the protein.An intron within the final exon is seen in Incyte template 1398043.12, which eliminates sequence from 2027-2079 in the mRNA, a region within the 3' UTR. There may also be another form with a longer intron in the last exon, eliminating the sequence from 2050-2584, which would cause a shift in reading frame, and open the reading frame until the end of the mRNA.

SGPr157, SEQ ID NO:33, SEQ ID NO:68
Genomic DNA source: Celera Assembly 5h contig 90000625988051
Homologs used for Genewise: gi_11427093, dbj_BAB22991.1, ref_NP_060705.1

SGPr154, SEQ ID NO:34 SEQ ID NO:69
Genomic DNA source: Public genomic contigs: gi_9798229, gi_9798027
Homologs used for Genewise: gb_AAK22721.1, pir_T38349

SGPr088, SEQ ID NO:35, SEQ ID NO:70
Genomic DNA source: Celera Assembly 5h contig 90000625988051
Homologs used for Genewise: dbj_BAB22991.1, ref_NP_060705.1, gi_7023108
Notes: Alternative splice forms are predicted by Incyte EST template 997089.29, public sequence gi_10440455, and Sugen-built clusters of public ESTs: cluster2209_-11_ncbi, cluster2209_-15_ncbi and cluster2209_-14_ncbi. All result in truncated proteins with short unique C-termini.

DESCRIPTION OF NOVEL PROTEASE POLYNUCLEOTIDES

SGPr140, SEQ ID NO:1, SEQ ID NO:36 is 1140 nucleotides long. The open reading frame starts at position 1 and ends at position 1140, giving an ORF length of 1140 nucleotides. The predicted protein is 379 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Aspartyl, PepsinA1. This gene maps to chromosomal position 1p13-p33. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):ctggtggggcctggy, ss2008313_allelePos=201 ; ctctgtctactgcaacagk, ss703383_allelePos=201. SNP ss2008313 occurs at nucleotide 846 (aa 282) of the ORF (C or T = Gly or Gly) (silent). SNP ss703383 occurs at nucleotide 321 (aa 107) of the ORF (G or T = Arg or Ser). This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:A969042, AA411567. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 295 tgggtgccctctgtctactgc 315.

SGPr197, SEQ ID NO:2, SEQ ID NO:37 is 1500 nucleotides long. The open reading frame starts at position 1 and ends at position 1500, giving an ORF length of 1500 nucleotides. The predicted protein is 499 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Aspartyl, PepsinA1. This gene maps to chromosomal position 6p21.1.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BF727344, BG394217, AW297327.

SGPr005, SEQ ID NO:3, SEQ ID NO:38 is 1173 nucleotides long. The open reading frame starts at position 1 and ends at position 1173, giving an ORF length of 1173 nucleotides. The predicted protein is 390 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Aspartyl, PepsinA1. This gene maps to chromosomal position 1p33.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs: none.

SGPr078, SEQ ID NO:4, SEQ ID NO:39 is 1239 nucleotides long. The open reading frame starts at position 1 and ends at position 1239, giving an ORF length of 1239 nucleotides. The predicted protein is 412 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Aspartyl, PepsinA1. This gene maps to chromosomal position 11p15. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):aagtactcccaggy, ss20182_allelePos=101 . SNP ss20182 occurs at nucleotide 173 (aa 58) of the ORF (C or T = Ala or Val).. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BG260401, BF025894, BF793219.

SGPr084, SEQ ID NO:5, SEQ ID NO:40 is 1191 nucleotides long. The open reading frame starts at position 1 and ends at position 1191, giving an ORF length of 1191 nucleotides. The predicted protein is 396 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, HH. This gene maps to chromosomal position 12q11.

SGPr009, SEQ ID NO:6, SEQ ID NO:41 is 1137 nucleotides long. The open reading frame starts at position 1 and ends at position 1137, giving an ORF length of 1137 nucleotides. The predicted protein is 378 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, ICEp10. This gene maps to chromosomal position 11q22 This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):tgatggaaaataatgtr, ss726380_allelePos=201; gagacagctcaaay, ss866796_allelePos=187. ss726380 occurs at nucleotide 102 (aa 34) of the ORF (G or A = Val or Val) (silent). SNP ss866796 occurs at nucleotide 200 (aa 67) of the ORF (C or T = Tyr or Ile).. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 900 cttcattgctttcaaatcttcc 921; 77 ttgatgatttgatggaaaat 96.

SGPr286, SEQ ID NO:7, SEQ ID NO:42 is 705 nucleotides long. The open reading frame starts at position 1 and ends at position 705, giving an ORF length of 705 nucleotides. The predicted protein is 234 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, ICEp20. This gene maps to chromosomal position na. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):ytatgtggcctatcgcgatg; rs551848_allelePos=3135. SNP rs551848 occurs at nucleotide 489 (aa 163) of the

ORF (C or T = Gly or Gly) silent.. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 574 ctggagctgctgactgagg 592; 388 gtggggcccacagctctcc 406.

SGPr008, SEQ ID NO:8, SEQ ID NO:43 is 2010 nucleotides long. The open reading frame starts at position 1 and ends at position 2010, giving an ORF length of 2010 nucleotides. The predicted protein is 669 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, PepC2. This gene maps to chromosomal position 2p23 This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):rccgaatggagagggcg, ss678494_allelePos=201 . SNP ss678494 occurs at nucleotide 838 (aa 280) of the ORF (G or A = Ala or Thr).. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BE075751.

SGPr198, SEQ ID NO:9, SEQ ID NO:44 is 2112 nucleotides long. The open reading frame starts at position 1 and ends at position 2112, giving an ORF length of 2112 nucleotides. The predicted protein is 703 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, PepC2. This gene maps to chromosomal position 1q42.11.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BE047777, AW339160.

SGPr210, SEQ ID NO:10, SEQ ID NO:45 is 2127 nucleotides long. The open reading frame starts at position 1 and ends at position 2127, giving an ORF length of 2127 nucleotides. The predicted protein is 708 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, PepC2. This gene maps to chromosomal position 19q13.2. This

nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):ggttccttgcagcy, ss1376193_allelePos=473 . SNP ss1376193 occurs at nucleotide 330 (aa 110) of the ORF (C or T = Ala or Ala) silent.. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BE872274. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 1180 gaggaggatgacgaggatgagg 1201.

SGPr290, SEQ ID NO:11, SEQ ID NO:46 is 2136 nucleotides long. The open reading frame starts at position 1 and ends at position 2136, giving an ORF length of 2136 nucleotides. The predicted protein is 711 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, PepC2. This gene maps to chromosomal position 2p23. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 1835 agcagctgcacgctgccatg 1854.

SGPr116, SEQ ID NO:12, SEQ ID NO:47 is 2109 nucleotides long. The open reading frame starts at position 1 and ends at position 2109, giving an ORF length of 2109 nucleotides. The predicted protein is 702 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, PepC2. This gene maps to chromosomal position 6p12. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 1003 ctggagatctgcaacctcac 1022.

SGPr003, SEQ ID NO:13, SEQ ID NO:48 is 1542 nucleotides long. The open reading frame starts at position 1 and ends at position 1542, giving an ORF length of 1542 nucleotides. The predicted protein is 513 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Cysteine, PepC2. This gene maps to chromosomal position 2q37.This

sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AL526645, BG475966, AL529373. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 1520 gctgctgcaggagccgctgctg 1541.

SGPr016, SEQ ID NO:14, SEQ ID NO:49 is 846 nucleotides long. The open reading frame starts at position 1 and ends at position 846, giving an ORF length of 846 nucleotides. The predicted protein is 281 amino acids long. This sequence codes for a partial protein.It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AW589885, AI024863. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 710 ttaaatatatttcttctcataa 731.

SGPr352, SEQ ID NO:15, SEQ ID NO:50 is 3312 nucleotides long. The open reading frame starts at position 1 and ends at position 3312, giving an ORF length of 3312 nucleotides. The predicted protein is 1103 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 19p13.3.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AW027573, AI131032, AI193804. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 1335 agactcgggcctggggctct 1354.

SGPr050, SEQ ID NO:16, SEQ ID NO:51 is 3675 nucleotides long. The open reading frame starts at position 1 and ends at position 3675, giving an ORF length of 3675 nucleotides. The predicted protein is 1224 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 5q15.3. This nucleotide sequence contains the following single nucleotide

polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):tcggctgaaaggcy, ss1483925_allelePos=216 . SNP ss1483925 occurs at nucleotide 310 (aa 104) of the ORF (C or T = Pro or Ser). This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BF933693. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 2067 tttcttcttttctttgtcaa 2086; 2061 atttgatttcttcttttctt 2080.

SGPr282, SEQ ID NO:17, SEQ ID NO:52 is 2196 nucleotides long. The open reading frame starts at position 1 and ends at position 2196, giving an ORF length of 2196 nucleotides. The predicted protein is 731 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 16p12.3. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):ggcaatataaaaggcy, ss679422_allelePos=201; acttcactgggctay, ss647742_allelePos=201; ggccgagcccaacgcaay, ss1226992_allelePos=101 . SNP ss679422 occurs at nucleotide 625 (aa 209) of the ORF (C or T = His or Tyr). SNP ss647742 occurs at nucleotide 1893 (aa 631) of the ORF (C or T = Tyr or Tyr) silent. SNP ss1226992 occurs at nucleotide 500 (aa 166) of the ORF (C or T = Thr or Met)..

SGPr046, SEQ ID NO:18, SEQ ID NO:53 is 2805 nucleotides long. The open reading frame starts at position 1 and ends at position 2805, giving an ORF length of 2805 nucleotides. The predicted protein is 934 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position

16q23The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 2353 gtgaggaagagggagatgaagt 2374.

SGPr060, SEQ ID NO:19, SEQ ID NO:54 is 4287 nucleotides long. The open reading frame starts at position 1 and ends at position 4287, giving an ORF length of 4287 nucleotides. The predicted protein is 1428 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 15q26.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AW575922, AW341169.

SGPr068, SEQ ID NO:20, SEQ ID NO:55 is 3561 nucleotides long. The open reading frame starts at position 1 and ends at position 3561, giving an ORF length of 3561 nucleotides. The predicted protein is 1186 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 10q22.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AJ403134, .

SGPr096, SEQ ID NO:21, SEQ ID NO:56 is 5808 nucleotides long. The open reading frame starts at position 1 and ends at position 5808, giving an ORF length of 5808 nucleotides. The predicted protein is 1935 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 3p14.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BE164543, AW995949, BF842288.

SGPr119, SEQ ID NO:22, SEQ ID NO:57 is 4518 nucleotides long. The open reading frame starts at position 1 and ends at position 4518, giving an ORF length of 4518 nucleotides. The predicted protein is 1505 amino acids long. This sequence

codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 12q11-q12.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AU132053, . The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 1257 taaagaaatgaaagttacaaa 1277.

SGPr143, SEQ ID NO:23, SEQ ID NO:58 is 2649 nucleotides long. The open reading frame starts at position 1 and ends at position 2649, giving an ORF length of 2649 nucleotides. The predicted protein is 882 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 20p13. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):ggcagtggctactgcy, ss787708_allelePos=201 . SNP ss787708 occurs at nucleotide 1750 (aa 584) of the ORF (C or T = Arg or Trp). . This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AA442551. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 2212 tgccactgtgctccaggctg 2231. This protein is predicted to have a transmembrane helix between amino acids 78 and 100. (TMHMM, a Hidden Markov Model based transmenbrane prediction program, Sonnhammer, et al Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology, p 175-182 AAAI Press, 1998.)

SGPr164, SEQ ID NO:24, SEQ ID NO:59 is 2937 nucleotides long. The open reading frame starts at position 1 and ends at position 2937, giving an ORF length of 2937 nucleotides. The predicted protein is 978 amino acids long. This sequence codes for a nearly full length protein with only the N terminus missing. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This

gene maps to chromosomal position 11q25. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):ataccgatcctgcaay, ss76755_allelePos=87 . SNP ss76755 occurs at nucleotide 1773 (aa 591) of the ORF (C or T = Asn or Asn) silent..

SGPr281, SEQ ID NO:25, SEQ ID NO:60 is 3285 nucleotides long. The open reading frame starts at position 1 and ends at position 3285, giving an ORF length of 3285 nucleotides. The predicted protein is 1094 amino acids long. This sequence codes for a nearly full length protein, with just the amino terminus missing.It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position 5q31.

SGPr075, SEQ ID NO:26, SEQ ID NO:61 is 375 nucleotides long. The open reading frame starts at position 1 and ends at position 375, giving an ORF length of 375 nucleotides. The predicted protein is 125 amino acids long. This sequence codes for a partial protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, ADAM. This gene maps to chromosomal position na.

SGPr292, SEQ ID NO:27, SEQ ID NO:62 is 1710 nucleotides long. The open reading frame starts at position 1 and ends at position 1710, giving an ORF length of 1710 nucleotides. The predicted protein is 569 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM10. This gene maps to chromosomal position 10q26.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AW665196. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 52 gctccctggcccacccagcc 71; 959 aagcaattcaaaagctgtatg 979.

SGPr069, SEQ ID NO:28, SEQ ID NO:63 is 2232 nucleotides long. The open reading frame starts at position 1 and ends at position 2232, giving an ORF length of 2232 nucleotides. The predicted protein is 743 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM13. This gene maps to chromosomal position Chr. SGPr212, SEQ ID NO:29, SEQ ID NO:64 is 2730 nucleotides long. The open reading frame starts at position 1 and ends at position 2730, giving an ORF length of 2730 nucleotides. The predicted protein is 909 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM1. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AL523882, T11456.

SGPr049, SEQ ID NO:30, SEQ ID NO:65 is 2973 nucleotides long. The open reading frame starts at position 1 and ends at position 2973, giving an ORF length of 2973 nucleotides. The predicted protein is 990 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM1. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AI222989. The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 2269 aatttaatatggaatatttat 2289.

SGPr026, SEQ ID NO:31, SEQ ID NO:66 is 1953 nucleotides long. The open reading frame starts at position 1 and ends at position 1953, giving an ORF length of 1953 nucleotides. The predicted protein is 650 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM1..

SGPr203, SEQ ID NO:32, SEQ ID NO:67 is 2175 nucleotides long. The open reading frame starts at position 1 and ends at position 2175, giving an ORF length of 2175 nucleotides. The predicted protein is 724 amino acids long. This sequence

codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM1. This gene maps to chromosomal position 2q37. This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AU132908, BE735172, BE563549 (many). The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 83 tggacgtggcctcggcctcca 103.

SGPr157, SEQ ID NO:33, SEQ ID NO:68 is 1524 nucleotides long. The open reading frame starts at position 1 and ends at position 1524, giving an ORF length of 1524 nucleotides. The predicted protein is 507 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM20. This gene maps to chromosomal position 18q22.3.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:BE386438, BE386547, BF920454 (many). The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 614 ccctggaggaacttgtggaa 633; 561 tcctgtgaatatcaaattca 580.

SGPr154, SEQ ID NO:34, SEQ ID NO:69 is 1422 nucleotides long. The open reading frame starts at position 1 and ends at position 1422, giving an ORF length of 1422 nucleotides. The predicted protein is 473 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM20. This nucleotide sequence contains the following single nucleotide polymorphisms (sequence preceding SNP is given, followed by identity of SNP, the accession number of SNP, and the allele position of SNP in the reference sequence):gtcatctatggty, ss1289877_allelePos=223. SNP ss1289877 occurs at nucleotide 457 (aa 153) of the ORF (C or T = Arg or Trp). The nucleic acid contains short repetitive sequence (the position and sequence of the repeat): 806 tccttgcagctgctgtcagc 825.

SGPr088, SEQ ID NO:35, SEQ ID NO:70 is 1428 nucleotides long. The open reading frame starts at position 1 and ends at position 1428, giving an ORF length of 1428 nucleotides. The predicted protein is 475 amino acids long. This sequence codes for a full length protein. It is classified as (superfamily/group/family): Protease, Metalloprotease, PepM20. This gene maps to chromosomal position 18q23.This sequence is represented in the database of public ESTs (dbEST) by the following ESTs:AL541127, AL542184, AL529661 (many).

EXAMPLE 2: Expression Analysis of Mammalian Proteases

Materials and Methods

Quantitative PCR Analysis

RNA is isolated from a variety of normal human tissues and cell lines. Single stranded cDNA is synthesized from 10 μg of each RNA as described above using the Superscript Preamplification System (GibcoBRL). These single strand templates are then linearly amplified with a pair of specific primers in a real time PCR reaction on a Light Cycler (Roche Molecular Biochemical). Graphical readout can provide quantitative analysis of the relative abundance of the targeted gene in the total RNA preparation.

DNA Array Based Expression Analysis

DNA-free RNA is isolated from a variety of normal human tissues, cryostat sections, and cell lines. Single stranded cDNA is synthesized from 10ug RNA or 1ug mRNA using a modification of the SMART PCR cDNA synthesis technique (Clontech). The procedure can be modified to allow asymmetric labeling of the 5′ and 3′ ends of each transcript with a unique oligonucleotide sequence. The resulting sscDNAs are then linearly amplified using Advantage long-range PCR (Clontech) on a Light Cycler PCR machine. Reactions are halted when the graphical real-time display demonstrates the products have begun to plateau. The double stranded cDNA products are purified using Millipore DNA purification matrix, dried,

resuspended, quantified, and analyzed on an agarose gel. The resulting elements are referred to as "tissue cDNAs".

Tissue cDNAs are spotted onto GAPS coated glass slides (Corning) using a Genetic Microsystems (GMS) arrayer at 500 ng/ul.

Fluorescent labeled oligonucleotides are synthesized to each novel exon, ensuring they contained internal mismatches with the closest known homologue. Typically oligos are 45 nucleotides long, labeled on the 5′ end with Cy5.

Exon-specific Cy5-labeled oligos are hybridized to the tissue cDNAs arrayed onto glass slides, and washed using standard buffers and conditions. Hybridizing signals are then quantified using a GMS Scanner.

Alternatively, tissue cDNAs are manually spotted onto Nylon membranes using a 384 pin replicator, and hybridized to ^{32}P-end labeled oligo probes.

Tissue cDNAs are generated from multiple RNA templates selected to provide information of relevance to the disease areas of interest and to reflect the biological mechanism of action for each protease. These templates include: human tumor cell lines, cryostat sections of primary human tumors and 32 normal human tissues to identify cancer-related genes; sections of normal, Alzheimer's, Parkinson's, and Schizophrenia brain regions for CNS-related genes; normal and diabetic or obese skeletal muscle, adipose, or liver for metabolic-related genes; and purified hematopoeitic cells, and lymphoid tissues for immune-related genes. To characterize gene mechanism of action, tissue cDNAs are generated to reflect angiogenesis (cultured endothelial cells treated with VEGF ligand, anti-angiogenic drugs, or hypoxia), motility (A549 cells stimulated with HGF ligand, orthotopic metastases, primary tumors with matched metastatic tumors), cell cycle (Hela, H1299, and other cell lines synchronized by drug block and harvested at various times in the cell cycle), checkpoint integrity and DNA repair (p53 normal or defective cells treated with γ-radiation, UV, cis-platinum, or oxidative stress), and cell survival (cells induced to differentiate or at various stages of apoptosis).

DESCRIPTION OF NOVEL PROTEASE POLYPEPTIDES

SGPr140, SEQ ID NO:1, SEQ ID NO:36 encodes a protein that is 379 amino acids long. It is classified as an Aspartylprotease, of the PepsinA1 family. The protease domain in this protein matches the hidden Markov profile for a Eukaryotic aspartyl protease, from amino acid 65 to amino acid 378. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 356. Other domains identified within this protein are: none. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.40E-160; number of identical amino acids = 263; percent identity = 66%; percent similarity = 76%; the accession number of the most similar entry in NRAA is CAC19554.1; the name or description, and species, of the most similar protein in NRAA is: Chymosin [Camelus dromedarius].

SGPr197, SEQ ID NO:2, SEQ ID NO:37 encodes a protein that is 499 amino acids long. It is classified as an Aspartylprotease, of the PepsinA1 family. The protease domain in this protein matches the hidden Markov profile for a Ubiquitin carboxyl-terminal hydrolases family 2, from amino acid 199 to amino acid 230. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 32. Other domains identified within this protein are: Zn-finger in ubiquitin-hydrolases (amino acid 26 to amino acid 96) P_Score = 5.6e-025. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 6.90E-137; number of identical amino acids = 296; percent identity = 46%; percent similarity = 56%; the accession number of the most similar entry in NRAA is CAB66759.1; the name or description, and species, of the most similar protein in NRAA is: Hypothetical histone deacetylase [Homo sapiens].

SGPr005, SEQ ID NO:3, SEQ ID NO:38 encodes a protein that is 390 amino acids long. It is classified as an Aspartylprotease, of the PepsinA1 family. The protease domain in this protein matches the hidden Markov profile for a Eukaryotic aspartyl protease, from amino acid 65 to amino acid 389. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 356. Other domains identified within this protein are: none. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.40E-130; number of identical amino acids = 230; percent identity = 62%; percent similarity = 76%; the accession number of the most similar entry in NRAA is BAB11755.1; the name or description, and species, of the most similar protein in NRAA is: Pepsinogen C [Rhinolophus ferrumequinum].

SGPr078, SEQ ID NO:4, SEQ ID NO:39 encodes a protein that is 412 amino acids long. It is classified as an Aspartylprotease, of the PepsinA1 family. The protease domain in this protein matches the hidden Markov profile for a Eukaryotic aspartyl protease, from amino acid 70 to amino acid 409. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 356. Other domains identified within this protein are: none. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 3.20E-285; number of identical amino acids = 412; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is NP_001900.1; the name or description, and species, of the most similar protein in NRAA is: Cathepsin D (lysosomal aspartyl protease) [Homo sapiens].

SGPr084, SEQ ID NO:5, SEQ ID NO:40 encodes a protein that is 396 amino acids long. It is classified as a Cysteineprotease, of the HH family. The protease domain in this protein matches the hidden Markov profile for a Hedgehog amino-terminal

signaling domain, from amino acid 23 to amino acid 185. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 163. Other domains identified within this protein are: Hint module amino acids 188-396; P_Score = 5.9e-120. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 3.00E-259; number of identical amino acids = 396; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is O43323; the name or description, and species, of the most similar protein in NRAA is: DESERT HEDGEHOG PROTEIN PRECURSOR (DHH) (HHG-3) [Homo sapiens].

SGPr009, SEQ ID NO:6, SEQ ID NO:41 encodes a protein that is 378 amino acids long. It is classified as a Cysteineprotease, of the ICEp10 family. The protease domain in this protein matches the hidden Markov profile for a ICE-like protease (caspase) p20 domain, from amino acid 131 to amino acid 264. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 141. Other domains identified within this protein are: ICE-like protease (caspase) p10 domain, amino acids 291-376; profile from 1-95: Caspase recruitment domain from amino acids 2-91. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 3.50E-129; number of identical amino acids = 223; percent identity = 55%; percent similarity = 67%; the accession number of the most similar entry in NRAA is NP_033938.1; the name or description, and species, of the most similar protein in NRAA is: Caspase 12 [Mus musculus].

SGPr286, SEQ ID NO:7, SEQ ID NO:42 encodes a protein that is 234 amino acids long. It is classified as a Cysteineprotease, of the ICEp20 family. The protease domain in this protein matches the hidden Markov profile for a ICE-like protease

(caspase) p20 domain, from amino acid 19 to amino acid 58. The positions within the HMMR profile that match the protein sequence are from profile position 22 to profile position 61. Other domains identified within this protein are: ICE-like protease (caspase) p10 domain, amino acids 144-202; profile from 1-61.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 4.60E-42; number of identical amino acids = 108; percent identity = 46%; percent similarity = 65%; the accession number of the most similar entry in NRAA is NP_036246.1; the name or description, and species, of the most similar protein in NRAA is: Caspase 14, apoptosis-related cysteine protease [Homo sapiens].

SGPr008, SEQ ID NO:8, SEQ ID NO:43 encodes a protein that is 669 amino acids long. It is classified as a Cysteineprotease, of the PepC2 family. The protease domain in this protein matches the hidden Markov profile for a Calpain family cysteine protease; Peptidase_C2, from amino acid 35 to amino acid 333. The positions within the HMMR profile that match the protein sequence are from profile position 2 to profile position 344. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 9.10E-86; number of identical amino acids = 229; percent identity = 33%; percent similarity = 53%; the accession number of the most similar entry in NRAA is AAD34601.1; the name or description, and species, of the most similar protein in NRAA is: Lens-specific calpain Lp82 [Oryctolagus cuniculus].

SGPr198, SEQ ID NO:9, SEQ ID NO:44 encodes a protein that is 703 amino acids long. It is classified as a Cysteineprotease, of the PepC2 family. The protease domain in this protein matches the hidden Markov profile for a Calpain family cysteine protease; Peptidase_C2, from amino acid 45 to amino acid 344. The positions within the HMMR profile that match the protein sequence are from profile

position 1 to profile position 344. Other domains identified within this protein are: Calpain large subunit, domain III, amino acids 355-512, profile from 1-163. Also three EF hand motifs at amino acids 579-607, 609-637 and 674-701; all EF hands match from 1-26 of profile. . The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 593; percent identity = 84%; percent similarity = 92%; the accession number of the most similar entry in NRAA is BAA03369.1; the name or description, and species, of the most similar protein in NRAA is: Calpain [Rattus norvegicus].

SGPr210, SEQ ID NO:10, SEQ ID NO:45 encodes a protein that is 708 amino acids long. It is classified as a Cysteineprotease, of the PepC2 family. The protease domain in this protein matches the hidden Markov profile for a Calpain family cysteine protease; Peptidase_C2, from amino acid 45 to amino acid 341. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 344. Other domains identified within this protein are: Calpain large subunit, domain III, amino acids 353-499, profile from 1-163. Also one EF hand motif at amino acids 613-641; EF hand matches from 1-26 of profile. . The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 569; percent identity = 79%; percent similarity = 86%; the accession number of the most similar entry in NRAA is CAC10066.1; the name or description, and species, of the most similar protein in NRAA is: Calpain 12 [Mus musculus].

SGPr290, SEQ ID NO:11, SEQ ID NO:46 encodes a protein that is 711 amino acids long. It is classified as a Cysteineprotease, of the PepC2 family. The protease domain in this protein matches the hidden Markov profile for a Calpain family cysteine protease; Peptidase_C2, from amino acid 43 to amino acid 346. The

positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 344. Other domains identified within this protein are: Calpain large subunit, domain III, amino acids 347-490, profile from 1-163. Also two EF hand motifs at amino acids 561-593 and 595-622; EF hands match from 1-26 of profile. . The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 6.20E-103; number of identical amino acids = 256; percent identity = 39%; percent similarity = 56%; the accession number of the most similar entry in NRAA is AAD34601.1; the name or description, and species, of the most similar protein in NRAA is: Lens-specific calpain Lp82 [Oryctolagus cuniculus].

SGPr116, SEQ ID NO:12, SEQ ID NO:47 encodes a protein that is 702 amino acids long. It is classified as a Cysteineprotease, of the PepC2 family. The protease domain in this protein matches the hidden Markov profile for a Calpain family cysteine protease; Peptidase_C2, from amino acid 42 to amino acid 341. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 344. Other domains identified within this protein are: Calpain large subunit, domain III, amino acids 352-510, profile from 1-163. Also two EF hand motifs at amino acids 577-605 and 607-635; EF hands match from 1-26 of profile. . The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 702; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is NP_008989.1; the name or description, and species, of the most similar protein in NRAA is: Calpain 11 [Homo sapiens].

SGPr003, SEQ ID NO:13, SEQ ID NO:48 encodes a protein that is 513 amino acids long. It is classified as a Cysteineprotease, of the PepC2 family. The protease

domain in this protein matches the hidden Markov profile for a Calpain family cysteine protease; Peptidase_C2, from amino acid 13 to amino acid 322. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 344. Other domains identified within this protein are: Calpain large subunit, domain III, amino acids 338-494, profile from 3-163.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 513; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is NP_075574.1; the name or description, and species, of the most similar protein in NRAA is: Calpain 10 [Homo sapiens].

SGPr016, SEQ ID NO:14, SEQ ID NO:49 encodes a protein that is 281 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin family propeptide, Pep_M12B_propep, from amino acid 58 to amino acid 175. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 119. Other domains identified within this protein are: none. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.30E-89; number of identical amino acids = 215; percent identity = 52%; percent similarity = 58%; the accession number of the most similar entry in NRAA is S47656; the name or description, and species, of the most similar protein in NRAA is: tMDC II (ADAM 5-like) protein - crab-eating macaque.

SGPr352, SEQ ID NO:15, SEQ ID NO:50 encodes a protein that is 1103 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 239 to amino acid 457. The positions

within the HMMR profile that match the protein sequence are from profile position 1 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, from amino acids 90-201, matching profile from 1-119. Also five Thrombospondin type 1 domains from 551-601, 829-884, 888-944, 946-1002, 1007-1057. All thrombospondin type 1 domains match profile from 1-54.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 1072; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is AAG35563.1; the name or description, and species, of the most similar protein in NRAA is: Zinc metalloendopeptidase [Homo sapiens].

SGPr050, SEQ ID NO:16, SEQ ID NO:51 encodes a protein that is 1224 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 292 to amino acid 495. The positions within the HMMR profile that match the protein sequence are from profile position 3 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide from 111-235, matching profile from 1-119. Also has five Thrombospondin type 1 domains from 590-640, 930-986, 990-1047, 1055-1101, 1128-1180.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 6.80E-149; number of identical amino acids = 385; percent identity = 37%; percent similarity = 53%; the accession number of the most similar entry in NRAA is AAG35563.1; the name or description, and species, of the most similar protein in NRAA is: Zinc metalloendopeptidase [Homo sapiens].

SGPr282, SEQ ID NO:17, SEQ ID NO:52 encodes a protein that is 731 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin family propeptide, Pep_M12B_propep, from amino acid 75 to amino acid 190. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 119. Other domains identified within this protein are: Disintegrin domain at amino acids 415-487; matches profile from 4-86. Also EGF-like domain at amino acids 633-661.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 619; percent identity = 85%; percent similarity = 91%; the accession number of the most similar entry in NRAA is I52361; the name or description, and species, of the most similar protein in NRAA is: Metalloproteinase-like, disintegrin-like, cysteine-rich protein IVa [crab-eating macaque].

SGPr046, SEQ ID NO:18, SEQ ID NO:53 encodes a protein that is 934 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 1 to amino acid 194. The positions within the HMMR profile that match the protein sequence are from profile position 13 to profile position 203. Other domains identified within this protein are: Six Thrombospondin type 1 domains at 289-339, 569-627, 634-687, 689-736, 769-828, 844-890.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.10E-162; number of identical amino acids = 320; percent identity = 39%; percent similarity = 56%; the accession number of the most similar entry in NRAA is AAG35563.1; the name or description, and species, of the most similar protein in NRAA is: Zinc metalloendopeptidase [Homo sapiens].

SGPr060, SEQ ID NO:19, SEQ ID NO:54 encodes a protein that is 1428 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 639 to amino acid 860. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, Pep_M12B_propep from amino acids 502-615. Matches profile from 1-119. Also has one thrombospondin type 1 domain from 954-1004, matching profile from 1-54.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 5.20E-87; number of identical amino acids = 250; percent identity = 39%; percent similarity = 55%; the accession number of the most similar entry in NRAA is NP_055087.1; the name or description, and species, of the most similar protein in NRAA is: Disintegrin-like and metalloprotease with thrombospondin type 1 motif, 7 [Homo sapiens].

SGPr068, SEQ ID NO:20, SEQ ID NO:55 encodes a protein that is 1186 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 261 to amino acid 460. The positions within the HMMR profile that match the protein sequence are from profile position 3 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, Pep_M12B_propep from amino acids 120-240, matching profile from 1-119. Also has four thrombospondin type 1 domains between 556 - 1021.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 624; percent identity = 64%; percent similarity =

77%; the accession number of the most similar entry in NRAA is O15072; the name or description, and species, of the most similar protein in NRAA is: ADAM-TS 3 PRECURSOR [Homo sapiens].

SGPr096, SEQ ID NO:21, SEQ ID NO:56 encodes a protein that is 1935 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 293 to amino acid 499. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, Pep_M12B_propep from amino acids 112-242, matching profile from 1-119. Also has 13 thrombospondin type 1 domains between 589 - 1733.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 1465; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is BAA92550.1; the name or description, and species, of the most similar protein in NRAA is: KIAA1312 protein [Homo sapiens].

SGPr119, SEQ ID NO:22, SEQ ID NO:57 encodes a protein that is 1505 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 259 to amino acid 467. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, Pep_M12B_propep from amino acids 92-215, matching profile from 1-119. Also has eight thrombospondin type 1 domains between 561 - 1416.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences

(NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 699; percent identity = 53%; percent similarity = 70%; the accession number of the most similar entry in NRAA is BAA92550.1; the name or description, and species, of the most similar protein in NRAA is: KIAA1312 (ADAMS 9-like) protein [Homo sapiens].

SGPr143, SEQ ID NO:23, SEQ ID NO:58 encodes a protein that is 882 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 275 to amino acid 478. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, Pep_M12B_propep from amino acids 145-263, matching profile from 1-119. Also has Disintegrin motif 495-570.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 726; percent identity = 99%; percent similarity = 99%; the accession number of the most similar entry in NRAA is CAC16509.2; the name or description, and species, of the most similar protein in NRAA is: Novel disintegrin and reprolysin metalloproteinase [Homo sapiens].

SGPr164, SEQ ID NO:24, SEQ ID NO:59 encodes a protein that is 978 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 243 to amino acid 452. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 203. Other domains identified within this protein are: Reprolysin family propeptide, Pep_M12B_propep from amino acids 92-206, matching profile from 1-119. Also has three Thrombospondin type 1 domains from

amino acids 545 to 978. Also has Glucose-6-phosphate dehydrogenase motif at 855-878.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.80E-264; number of identical amino acids = 465; percent identity = 50%; percent similarity = 67%; the accession number of the most similar entry in NRAA is XP_012978.1; the name or description, and species, of the most similar protein in NRAA is: ADAMS-1 preproprotein [Homo sapiens].

SGPr281, SEQ ID NO:25, SEQ ID NO:60 encodes a protein that is 1094 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 317 to amino acid 432. The positions within the HMMR profile that match the protein sequence are from profile position 89 to profile position 203. Other domains identified within this protein are: Six Thrombospondin type 1 domains from amino acid 346 to 1030.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 4.4e-075; number of identical amino acids = 287; percent identity = 39%; percent similarity = 55%; the accession number of the most similar entry in NRAA is NP_112217.1; the name or description, and species, of the most similar protein in NRAA is: ADAMTS 12 [Homo sapiens].

SGPr075, SEQ ID NO:26, SEQ ID NO:61 encodes a protein that is 125 amino acids long. It is classified as a Metalloprotease, of the ADAM family. The protease domain in this protein matches the hidden Markov profile for a Reprolysin (M12B) family zinc metalloprotease, from amino acid 1 to amino acid 123. The positions within the HMMR profile that match the protein sequence are from profile position 14 to profile position 203. Other domains identified within this protein are: none. The results of a Smith Waterman search (PAM100, gap open and extend penalties of

12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.10E-54; number of identical amino acids = 98; percent identity = 65%; percent similarity = 73%; the accession number of the most similar entry in NRAA is CAC18729; the name or description, and species, of the most similar protein in NRAA is: Metalloprotease/disintegrin [Rattus norvegicus].

SGPr292, SEQ ID NO:27, SEQ ID NO:62 encodes a protein that is 569 amino acids long. It is classified as a Metalloprotease, of the PepM10 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase_M10, Matrixin, from amino acid 56 to amino acid 267. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 171. Other domains identified within this protein are: Also has four Hemopexin domains at amino acids 333-391, 394-449, 451-499, 506-549.. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 6.00E-137; number of identical amino acids = 333; percent identity = 57%; percent similarity = 74%; the accession number of the most similar entry in NRAA is AAC21447.1; the name or description, and species, of the most similar protein in NRAA is: Matrix metalloproteinase [Xenopus laevis].

SGPr069, SEQ ID NO:28, SEQ ID NO:63 encodes a protein that is 743 amino acids long. It is classified as a Metalloprotease, of the PepM13 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase family M13, from amino acid 535 to amino acid 742. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 225. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded

the following results: Pscore = 0; number of identical amino acids = 581; percent identity = 78%; percent similarity = 90%; the accession number of the most similar entry in NRAA is AAG18446.1; the name or description, and species, of the most similar protein in NRAA is: Neprilysin-like peptidase alpha [Mus musculus]. This protein is predicted to have a transmembrane helix between amino acids 13 and 35. This transmembrane region could function as a signal peptide. (TMHMM, a Hidden Markov Model based transmenbrane prediction program, Sonnhammer, et al Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology, p 175-182 AAAI Press, 1998.)

SGPr212, SEQ ID NO:29, SEQ ID NO:64 encodes a protein that is 909 amino acids long. It is classified as a Metalloprotease, of the PepM1 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase family M1, from amino acid 275 to amino acid 306. The positions within the HMMR profile that match the protein sequence are from profile position 343 to profile position 374. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.40E-31; number of identical amino acids = 55; percent identity = 77%; percent similarity = 87%; the accession number of the most similar entry in NRAA is BAB25647.1; the name or description, and species, of the most similar protein in NRAA is: Probable zinc metal proteinase [Mus musculus].

SGPr049, SEQ ID NO:30, SEQ ID NO:65 encodes a protein that is 990 amino acids long. It is classified as a Metalloprotease, of the PepM1 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase family M1, from amino acid 98 to amino acid 506. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 441. Other domains identified within this protein are: None. The results of a Smith

Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 4.10E-220; number of identical amino acids = 375; percent identity = 68%; percent similarity = 79%; the accession number of the most similar entry in NRAA is BAB29490.1; the name or description, and species, of the most similar protein in NRAA is: Putative aminopeptidase [Mus musculus]. This protein is predicted to have a transmembrane helix between amino acids 13 and 35. This transmembrane region could function as a signal peptide. (TMHMM, a Hidden Markov Model based transmenbrane prediction program, Sonnhammer, et al Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology, p 175-182 AAAI Press, 1998)

SGPr026, SEQ ID NO:31, SEQ ID NO:66 encodes a protein that is 650 amino acids long. It is classified as a Metalloprotease, of the PepM1 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase family M1, from amino acid 32 to amino acid 417. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 441. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 0; number of identical amino acids = 650; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is AAH01064; the name or description, and species, of the most similar protein in NRAA is: Hypothetical protein DKFZp547H084 [Homo sapiens].

SGPr203, SEQ ID NO:32, SEQ ID NO:67 encodes a protein that is 724 amino acids long. It is classified as a Metalloprotease, of the PepM1 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase family M1, from amino acid 194 to amino acid 444. The positions within the HMMR

profile that match the protein sequence are from profile position 161 to profile position 441. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.90E-276; number of identical amino acids = 493; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is AAG22080.1; the name or description, and species, of the most similar protein in NRAA is: RNPEP-like protein [Homo sapiens].

SGPr157, SEQ ID NO:33, SEQ ID NO:68 encodes a protein that is 507 amino acids long. It is classified as a Metalloprotease, of the PepM20 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase_M20 , from amino acid 106 to amino acid 450. The positions within the HMMR profile that match the protein sequence are from profile position 42 to profile position 368. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 7.50E-202; number of identical amino acids = 310; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is AAH04271.1; the name or description, and species, of the most similar protein in NRAA is: Hypothetical protein [Homo sapiens].

SGPr154, SEQ ID NO:34, SEQ ID NO:69 encodes a protein that is 473 amino acids long. It is classified as a Metalloprotease of the PepM20 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase_M20, from amino acid 55 to amino acid 286. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 247. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database

of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 1.90E-28; number of identical amino acids = 122; percent identity = 31%; percent similarity = 48%; the accession number of the most similar entry in NRAA is AAK22721.1; the name or description, and species, of the most similar protein in NRAA is: M20/M25/M40 family peptidase [Caulobacter crescentus].

SGPr088, SEQ ID NO:35, SEQ ID NO:70 encodes a protein that is 475 amino acids long. It is classified as a Metalloprotease of the PepM20 family. The protease domain in this protein matches the hidden Markov profile for a Peptidase_M20, from amino acid 22 to amino acid 417. The positions within the HMMR profile that match the protein sequence are from profile position 1 to profile position 368. Other domains identified within this protein are: None. The results of a Smith Waterman search (PAM100, gap open and extend penalties of 12 and 2) of the public database of amino acid sequences (NRAA) with this protein sequence yielded the following results: Pscore = 9.8e-315; number of identical amino acids = 475; percent identity = 100%; percent similarity = 100%; the accession number of the most similar entry in NRAA is XP_008819.1; the name or description, and species, of the most similar protein in NRAA is: Hypothetical protein FLJ10830 [Homo sapiens].

EXAMPLE 3: Isolation of cDNAs Encoding Mammalian Proteases

Materials and Methods

Identification of novel clones

Total RNAs are isolated using the Guanidine Salts/Phenol extraction protocol of Chomczynski and Sacchi (P. Chomczynski and N. Sacchi, *Anal. Biochem.* 162:156 (1987)) from primary human tumors, normal and tumor cell lines, normal human tissues, and sorted human hematopoietic cells. These RNAs are used to generate single-stranded cDNA using the Superscript Preamplification System (GIBCO BRL, Gaithersburg, MD; Gerard, GF *et al.* (1989), FOCUS 11, 66) under conditions recommended by the manufacturer. A typical reaction uses 10 μg total

RNA with 1.5 μg oligo(dT)$_{12-18}$ in a reaction volume of 60 μL. The product is treated with RNaseH and diluted to 100 μL with H_20. For subsequent PCR amplification, 1-4 μL of this sscDNA is used in each reaction.

Degenerate oligonucleotides are synthesized on an Applied Biosystems 3948 DNA synthesizer using established phosphoramidite chemistry, precipitated with ethanol and used unpurified for PCR. These primers are derived from the sense and antisense strands of conserved motifs within the catalytic domain of several proteases. Degenerate nucleotide residue designations are: N = A, C, G, or T; R = A or G; Y = C or T; H = A, C or T not G; D = A, G or T not C; S = C or G; and W = A or T.

PCR reactions are performed using degenerate primers applied to multiple single-stranded cDNAs. The primers are added at a final concentration of 5 μM each to a mixture containing 10 mM TrisHCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$, 200 μM each deoxynucleoside triphosphate, 0.001% gelatin, 1.5 U AmpliTaq DNA Polymerase (Perkin-Elmer/Cetus), and 1-4 μL cDNA. Following 3 min denaturation at 95 °C, the cycling conditions are 94 °C for 30 s, 50 °C for 1 min, and 72 °C for 1 min 45 s for 35 cycles. PCR fragments migrating between 300-350 bp are isolated from 2% agarose gels using the GeneClean Kit (Bio101), and T-A cloned into the pCRII vector (Invitrogen Corp. U.S.A.) according to the manufacturer's protocol.

Colonies are selected for mini plasmid DNA-preparations using Qiagen columns and the plasmid DNA is sequenced using a cycle sequencing dye-terminator kit with AmpliTaq DNA Polymerase, FS (ABI, Foster City, CA). Sequencing reaction products are run on an ABI Prism 377 DNA Sequencer, and analyzed using the BLAST alignment algorithm (Altschul, S.F. *et al.*, *J.Mol.Biol.* 215: 403-10).

Additional PCR strategies are employed to connect various PCR fragments or ESTs using exact or near exact oligonucleotide primers. PCR conditions are as described above except the annealing temperatures are calculated for each oligo pair using the formula: Tm = 4(G+C)+2(A+T).

Isolation of cDNA clones:

Human cDNA libraries are probed with PCR or EST fragments corresponding to protease-related genes. Probes are ^{32}P-labeled by random priming and used at $2x10^6$ cpm/mL following standard techniques for library screening. Pre-hybridization (3 h) and hybridization (overnight) are conducted at 42 °C in 5X SSC, 5X Denhart's solution, 2.5% dextran sulfate, 50 mM $Na_2PO_4/NaHPO_4$, pH 7.0, 50% formamide with 100 mg/mL denatured salmon sperm DNA. Stringent washes are performed at 65 °C in 0.1X SSC and 0.1% SDS. DNA sequencing is carried out on both strands using a cycle sequencing dye-terminator kit with AmpliTaq DNA Polymerase, FS (ABI, Foster City, CA). Sequencing reaction products are run on an ABI Prism 377 DNA Sequencer.

EXAMPLE 4: Expression Analysis of Mammalian Proteases

Materials and Methods

Northern blot analysis

Northern blots are prepared by running 10 μg total RNA isolated from 60 human tumor cell lines (such as HOP-92, EKVX, NCI-H23, NCI-H226, NCI-H322M, NCI-H460, NCI-H522, A549, HOP-62, OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, IGROV1, SK-OV-3, SNB-19, SNB-75, U251, SF-268, SF-295, SF-539, CCRF-CEM, K-562, MOLT-4, HL-60, RPMI 8226, SR, DU-145, PC-3, HT-29, HCC-2998, HCT-116, SW620, Colo 205, HTC15, KM-12, UO-31, SN12C, A498, CaKi1, RXF-393, ACHN, 786-0, TK-10, LOX IMVI, Malme-3M, SK-MEL-2, SK-MEL-5, SK-MEL-28, UACC-62, UACC-257, M14, MCF-7, MCF-7/ADR RES, Hs578T, MDA-MB-231, MDA-MB-435, MDA-N, BT-549, T47D), from human adult tissues (such as thymus, lung, duodenum, colon, testis, brain, cerebellum, cortex, salivary gland, liver, pancreas, kidney, spleen, stomach, uterus, prostate, skeletal muscle, placenta, mammary gland, bladder, lymph node, adipose tissue), and 2 human fetal normal tissues (fetal liver, fetal brain), on a denaturing formaldehyde 1.2% agarose gel and transferring to nylon membranes.

Filters are hybridized with random primed [α^{32}P]dCTP-labeled probes synthesized from the inserts of several of the protease genes. Hybridization is performed at 42 °C overnight in 6X SSC, 0.1% SDS, 1X Denhardt's solution, 100 μg/mL denatured herring sperm DNA with 1-2 x 10^6 cpm/mL of ^{32}P-labeled DNA probes. The filters are washed in 0.1X SSC/0.1% SDS, 65 °C, and exposed on a Molecular Dynamics phosphorimager.

Quantitative PCR analysis

RNA is isolated from a variety of normal human tissues and cell lines. Single stranded cDNA is synthesized from 10 μg of each RNA as described above using the Superscript Preamplification System (GibcoBRL). These single strand templates are then used in a 25 cycle PCR reaction with primers specific to each clone. Reaction products are electrophoresed on 2% agarose gels, stained with ethidium bromide and photographed on a UV light box. The relative intensity of the STK-specific bands were estimated for each sample.

DNA Array Based Expression Analysis

Plasmid DNA array blots are prepared by loading 0.5 μg denatured plasmid for each protease on a nylon membrane. The [γ^{32}P]dCTP labeled single stranded DNA probes are synthesized from the total RNA isolated from several human immune tissue sources or tumor cells (such as thymus, dendrocytes, mast cells, monocytes, B cells (primary, Jurkat, RPMI8226, SR), T cells (CD8/CD4+, TH1, TH2, CEM, MOLT4), K562 (megakaryocytes). Hybridization is performed at 42 °C for 16 hours in 6X SSC, 0.1% SDS, 1X Denhardt's solution, 100 μg/mL denatured herring sperm DNA with 10^6 cpm/mL of [γ^{32}P]dCTP labeled single stranded probe. The filters are washed in 0.1X SSC/0.1% SDS, 65 °C, and exposed for quantitative analysis on a Molecular Dynamics phosphorimager.

EXAMPLE 5: Protease Gene Expression

Vector Construction

Materials and Methods

Expression Vector Construction

Expression constructs are generated for some of the human cDNAs including: a) full-length clones in a pCDNA expression vector; and b) a GST-fusion construct containing the catalytic domain of the novel protease fused to the C-terminal end of a GST expression cassette; and c) a full-length clone containing a mutation within the predicted polypeptide cleaving site within the protease domain, inserted in the pCDNA vector.

These mutants of the protease might function as dominant negative constructs, and will be used to elucidate the function of these novel proteases.

EXAMPLE 6: Generation of Specific Immunoreagents to Proteases

Materials and Methods

Specific immunoreagents are raised in rabbits against KLH- or MAP-conjugated synthetic peptides corresponding to isolated protease polypeptides. C-terminal peptides were conjugated to KLH with glutaraldehyde, leaving a free C-terminus. Internal peptides were MAP-conjugated with a blocked N-terminus. Additional immunoreagents can also be generated by immunizing rabbits with the bacterially expressed GST-fusion proteins containing the cytoplasmic domains of each novel PTK or STK.

The various immune sera are first tested for reactivity and selectivity to recombinant protein, prior to testing for endogenous sources.

Western blots

Proteins in SDS PAGE are transferred to immobilon membrane. The washing buffer is PBST (standard phosphate-buffered saline pH 7.4 + 0.1% Triton X-100). Blocking and antibody incubation buffer is PBST +5% milk. Antibody dilutions are varied from 1:1000 to 1:2000.

EXAMPLE 7: Recombinant Expression and Biological Assays for Proteases

Materials and Methods

Transient Expression of Proteases in Mammalian Cells

The pcDNA expression plasmids (10 μg DNA/100 mm plate) containing the protease constructs are introduced into 293 cells with lipofectamine (Gibco BRL). After 72 hours, the cells are harvested in 0.5 mL solubilization buffer (20 mM HEPES, pH 7.35, 150 mM NaCl, 10% glycerol, 1% Triton X-100, 1.5 mM $MgCl_2$, 1 mM EGTA, 2 mM phenylmethylsulfonyl fluoride, 1 μg/mL aprotinin). Sample aliquots are resolved by SDS polyacrylamide gel electrophoresis (PAGE) on 6% acrylamide/0.5% bis-acrylamide gels and electrophoretically transferred to nitrocellulose. Non-specific binding is blocked by preincubating blots in Blotto (phosphate buffered saline containing 5% w/v non-fat dried milk and 0.2% v/v nonidet P-40 (Sigma)), and recombinant protein is detected using the various anti-peptide or anti-GST-fusion specific antisera.

In Vitro Protease Assays

In vitro Protease Assay Using Fluorogenic Peptides

Assays are carried out using a spectrofluorometer, such as Perkin-Elmer 204S. The standard reaction mixtures (100 μl) contains 200 mM Tris-HCl, pH8.5, and 200 μM fluorogenic peptide substrate. After enzyme addition, reaction mixtures are incubated at 37 °C for 30 min and terminated by addition of 1.9 ml of 125 mM ZnSO4 (Brenner, C., and Fuller, R. S., 1992, *Proc. Natl. Acad. Sci. U. S. A.* 89:922-926). The precipitate is removed by centrifugation for 1 min in a microcentrifuge (15,000 × g), and the rate of product (7-amino-4-methyl-coumarin) released into the supernatant solution is determined fluorometrically [(excitation) = 385 nm, (emission) = 465 nm]. Examples of substrates used in the literature include: Boc-Gly-Arg-Arg-4-methylcoumaryl-7-amide (MCA), Boc-Gln-Arg-Arg-MCA, Z-Arg-Arg-MCA, and pGlu-Arg-Thr-Lys-Arg-MCA. Stock solutions (100 mM) are prepared by dissolving peptides in dimethyl sulfoxide that are then diluted in water

to 1 mM working stock before use. (Details of this assay can be found in: R. Yosuf, *et al. J. Biol. Chem.*, Vol. 275, Issue 14, 9963-9969, April 7, 2000 which is incorporated herein by reference in its entirety including any figures, tables, or drawings.)

Protease assay in intact cells using fluorogenic peptides-

Calpain activity is measured by the rate of generation of the fluorescent product, AMC, from intracellular thiol-conjugated Boc-Leu-Met-CMAC (Rosser, B. G., Powers, S. P., and Gores, G. J. (1993) *J. Biol. Chem.* 268, 23593-23600). Cells are dispersed, grown on glass coverslips, continuously superfused with physiologic saline solution at 37 °C, and sequentially imaged with a quantitative fluorescence imaging system. At t = 0, Boc-Leu-Met-CMAC (10 μM, Molecular Probes) is introduced into the superfusion solution, and mean fluorescence intensity (excitation 350 nm, emission 470 nm) of individual cells is measured at 60-s intervals. At 10 min, TNF- (30 ng/ml) is added to the superfusion solution with 10 μM Boc-Leu-Met-CMAC. The slope of the fluorescence change with respect to time represents the intracellular calpain activity (Rosser, *et al.*, 1993, *J. Biol. Chem.* 268:23593-23600). For calpain assays in whole cell populations, suspension cultures of cells are loaded with 10 μM Boc-Leu-Met-CMAC, and changes in intracellular fluorescence are measured prior to and after TNFalpha addition at 37 °C using a FACS Vantage system. Cellular fluorescence of AMC is measured using a 360-nm excitation filter and a 405-nm long-pass emission filter. (Details of this assay can be found in: Han, *et al.*, 1999, *J Biol Chem*, 274:787-794 which is incorporated herein by reference in its entirety including any figures, tables, or drawings)

Protease assay using chromogenic substrates

The proteolytic activity of enzymes is measured using a commercially available assay system (Athena Environmental Sciences, Inc.). The assay employs a universal substrate of a dye-protein conjugate cross linked to a matrix. Protease activity is determined spectrophotometrically by measuring the absorbance of the

dye released from the matrix to the supernatant. Reaction vials containing the enzyme and substrate are incubated for 3 h at 37 °C. The activity is measured at different incubation times, and reactions are terminated by adding 500 µl of 0.2 N NaOH to each vial. The absorbance of the supernatant in each reaction vial is measured at 450 nm. The proteolytic activity is monitored using 10 µl (approximately 10 µg) of purified protein incubated with 5 µg of -casein (Sigma) in 50 mM Tris-HCl (pH 7.5) for 30 min, 1 h or 2 h at 37 °C. The reaction products are resolved by SDS-polyacrylamide gel electrophoresis and proteins visualized by staining with Coomassie Blue (Details of this assay can be found in: Faccio, *et al.*, 2000, *J Biol Chem*, 275:2581-2588 which is incorporated herein by reference in its entirety including any figures, tables, or drawings).

Protease assay using radiolabeled substrate bound to membranes-

Unlabeled protease is mixed with radiolabeled substrate-containing membranes in buffer (100 mM HEPES, 100 mM NaCl, 125 µM magnesium acetate, 125 µM zinc acetate, pH 7.5) and incubated at 30 °C. Typically, each reaction had a final volume of 80-100 µl. Each reaction is normalized to the same final concentration of lysis buffer components (25 mM Tris, 0.1 M sorbitol, 0.5 mM EDTA, 0.01% NaN_3, pH 7.5) because the amount of membranes added to each reaction is varied. To examine metal ion specificity, reactions are assembled without substrate and pretreated with 1.125 mM 1,10-orthophenanthroline for 20 min on ice. Subsequently, metal ions and substrate-containing membranes are added, and reactions are initiated by incubation at 30 °C; the additions result in dilution of the 1,10-orthophenanthroline to a final concentration of 1 mM. The metal ions are added in the form of acetate salts from 25-100 mM stock solutions (Zn^{2+}, Mg^{2+}, Cu^{2+}, Co^{2+}, or Ca^{2+}) that are first acidified with 2 mM concentrated HCl and then neutralized with 1 mM HEPES, pH 7.5; this step is necessary to achieve full solubilization of zinc acetate. For analysis by immunoprecipitation, samples are diluted 10-20× with immunoprecipitation buffer (Berkower, C., and Michaelis, S. (1991) *EMBO J.* 10:3777-3785) containing 0.1% SDS, cleared of insoluble material

(13,000 × g for 5-10 min at 4 °C), and immunoprecipitated with substrate-specific antibody. Alternatively, samples are solubilized by SDS (final concentration, 0.5%), boiled for 3 min, and directly immunoprecipitated after dilution with immunoprecipitation buffer. Immunoprecipitates are subjected to SDS-polyacrylamide gel electrophoresis as described, fixed for 7 min with 20% trichloroacetic acid, dried, and exposed to a PhosphorImager screen for detection and quantitation (Molecular Dynamics, Sunnyvale, CA). All of the above reagents can be purchased from Sigma. (Details of this assay can be found in: Schmidt, *et al.*, 2000, *J Biol Chem*, 275:6227-6233 which is incorporated herein by reference in its entirety including any figures, tables, or drawings). Variation of this assay to apply to substrate not bound to membrane is straightforward.

A comprehensive discussion of various protease assays can be found in: The Handbook of Proteolytic Enzymes by Alan J. Barrett (Editor), Neil D. Rawlings (Editor), J. Fred Woessner (Editor) (February 1998) Academic Press, San Diego; ISBN: 0-12-079370-9 (which is incorporated herein by reference in its entirety including any figures, tables, or drawings).

Similar assays are performed on bacterially expressed GST-fusion constructs of the proteases.

EXAMPLE 8a: Chromosomal Localization of Proteases

Materials And Methods

Several sources were used to find information about the chromosomal localization of each of the genes described in this patent application. First, cytogenetic map locations of these contigs were found in the title or text of their Genbank record, or by inspection through the NCBI human genome map viewer (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch?). Alternatively, the accession number of a genomic contig (identified by BLAST against NRNA) was used to query the Entrez Genome Browser (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/MapViewerHelp.html), and the

cytogenetic localization was read from the NCBI data. A thorough search of available literature for the cytogenetic region is also made using Medline (http://www.ncbi.nlm.nih.gov/PubMed/medline.html). References for association of the mapped sites with chromosomal amplifications found in human cancer can be found in: Knuutila, et al., Am J Pathol, 1998, 152:1107-1123.

Results

The chromosomal regions for mapped genes are listed Table 2. The chromosomal positions were cross-checked with the Online Mendelian Inheritance in Man database (OMIM, http://www.ncbi.nlm.nih.gov/htbin-post/Omim)., which tracks genetic information for many human diseases, including cancer. References for association of the mapped sites with chromosomal abnormalities found in human cancer can be found in: Knuutila, et al., Am J Pathol, 1998, 152:1107-1123. A third source of information on mapped positions was searching published literature (at NCBI, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) for documented association of the mapped position with human disease.

The following section describes various diseases that map to chromosomal locations established for proteases included in this patent application. The protease polynucleotides of the present invention can be used to identify individuals who have, or are at risk for developing, relevant diseases. As discussed elsewhere in this application, the polypeptides and polynucleotides of the present invention are useful in identifying compounds that modulate protease activity, and in turn ameliorate various diseases.

SGPr140 SEQ ID NO:1 1p33/1p13.3

Novel recurrent genetic imbalances in human hepatocellular carcinoma cell lines identified by comparative genomic hybridization. (Hepatology. 1999 Apr;29(4):1208-14.) Chromosome 1 alterations in breast cancer: allelic loss on 1p

and 1q is related to lymphogenic metastases and poor prognosis. (Genes Chromosomes Cancer. 1992 Nov;5(4):311-20.).

SGPr197 SEQ ID NO:2 6p21.1

Genetic imbalances with impact on survival in head and neck cancer patients. (Am J Pathol. 2000 Aug;157(2):369-75.). Systematic screening of the LDL-PLA2 gene for polymorphic variants and case-control analysis in schizophrenia. (Biochem Biophys Res Commun. 1997 Dec 29;241(3):630-5.)

SGPr005 SEQ ID NO:3 1p33

Novel recurrent genetic imbalances in human hepatocellular carcinoma cell lines identified by comparative genomic hybridization. (Hepatology. 1999 Apr;29(4):1208-14.) Chromosome 1 alterations in breast cancer: allelic loss on 1p and 1q is related to lymphogenic metastases and poor prognosis. (Genes Chromosomes Cancer. 1992 Nov;5(4):311-20.).

SGPr078 SEQ ID NO:4 11p15

Use of horizontal ultrathin gel electrophoresis to analyze allelic deletions in chromosome band 11p15.5 in gliomas. (Neuro-oncol. 2000 Jan;2(1):1-5.). Loss of heterozygosity and heterogeneity of its appearance and persisting in the course of acute myeloid leukemia and myelodysplastic syndromes. (Leuk Res. 2001 Jan;25(1):45-53.). Chromosomal localization of two genes underlying late-infantile neuronal ceroid lipofuscinosis. (Neurogenetics. 1998 Mar;1(3):217-22.). The usher syndromes also map to this location. (Am J Med Genet. 1999 Sep 24;89(3):158-66.)

SGPr084.2 SEQ ID NO:5 12q11

Fine genetic mapping of diffuse non-epidermolytic palmoplantar keratoderma to chromosome 12q11-q13: exclusion of the mapped type II keratins. (Exp Dermatol. 1999 Oct;8(5):388-91.).

SGPr009 SEQ ID NO:6 11q22

Restricted chromosome breakpoint sites on 11q22-q23.1 and 11q25 in various hematological malignancies without MLL/ALL-1 gene rearrangement. (Cancer Genet Cytogenet. 2001 Jan 1;124(1):27-35.). Molecular characterization of deletion at 11q22.1-23.3 in mantle cell lymphoma. (Br J Haematol. 1999 Mar;104(4):665-71.). Structure and chromosome localization of the human CASP8 gene (implicated in tumorigenesis, with loss of heterogeneity (LOH)). (Gene. 1999 Jan 21;226(2):225-32. Reduced expression of adhesion molecules and cell signaling receptors by chronic lymphocytic leukemia cells with 11q deletion. (Blood. 1999 Jan 15;93(2):624-31.).

SGPr286 SEQ ID NO:7 16p13.3

Monosomy for the most telomeric, gene-rich region of the short arm of human chromosome 16 causes minimal phenotypic effects (Eur J Hum Genet. 2001 Mar;9(3):217-225.). Identification of a subtle t(16;19)(p13.3;p13.3) in an infant with multiple congenital abnormalities using a 12-colour multiplex FISH telomere assay, M-TEL. (Eur J Hum Genet. 2000 Dec;8(12):903-10). Familial Mediterranean fever in the 'Chuetas' of Mallorca: a question of Jewish origin or genetic heterogeneity (Eur J Hum Genet. 2000 Apr;8(4):242-6.). Familial mental retardation syndrome ATR-16 due to an inherited cryptic subtelomeric translocation, t(3;16)(q29;p13.3) (Am J Hum Genet. 2000 Jan;66(1):16-25). Autosomal dominant polycystic kidney disease: clues to pathogenesis. (Hum Mol Genet. 1999;8(10):1861-6. Review).

SGPr008 SEQ ID NO:8 2p23

Familial syndromic esophageal atresia (Am J Hum Genet. 2000 Feb;66(2):436-44.). Chromosomal rearrangements in acute myelogenous leukemia involving loci on chromosome (Leukemia. 1999 Oct;13(10):1534-8.). Association and linkage analysis of candidate chromosomal regions in multiple sclerosis: indication of

disease genes in disease genes in 12q23 and 7ptr-15 (Eur J Hum Genet. 1999 Feb-Mar;7(2):110-6.).

SGPr198 SEQ ID NO:9 1q42

Familial effort polymorphic ventricular arrhythmias in arrhythmogenic right ventricular cardiomyopathy map to chromosome 1q42-43 (Am J Cardiol. 2000 Mar 1;85(5):573-9.). Replication linkage study for prostate cancer susceptibility genes (Prostate. 2000 Oct 1;45(2):106-14.). Linkage analyses at the chromosome 1 loci 1q24-25 (HPC1), 1q42.2-43 (PCAP), and 1p36 (CAPB) in families with hereditary prostate cancer (Am J Hum Genet. 2000 Feb;66(2):539-46.). Clinical profile and long-term follow-up of 37 families with arrhythmogenic right ventricular cardiomyopathy

(J Am Coll Cardiol. 2000 Dec;36(7):2226-33.) Arrhythmic disorder mapped to chromosome 1q42-q43 causes malignant polymorphic ventricular tachycardia in structurally normal hearts (J Am Coll Cardiol. 1999 Dec;34(7):2035-42.). Analysis of chromosome 1q42.2-43 in 152 families with high risk of prostate cancer. (Am J Hum Genet. 1999 Apr;64(4):1087-95.). A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families (Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14875-9.).

SGPr210 SEQ ID NO:10 19q13.2

A microdeletion in 19q13.2 associated with mental retardation, skeletal malformations, and Diamond-Blackfan anaemia suggests a novel contiguous gene syndrome (J Med Genet. 2000 Feb;37(2):128-31.). A microdeletion syndrome due to a 3-Mb deletion on 19q13.2--Diamond-Blackfan anemia associated with macrocephaly, hypotonia, and psychomotor retardation. (Clin Genet. 1999 Jun;55(6):487-92.). Diamond-Blackfan Anaemia: an overview. (Paediatr Drugs. 2000 Sep-Oct;2(5):345-55. Review.) A microdeletion in 19q13.2 associated with mental retardation, skeletal malformations, and Diamond-Blackfan anaemia suggests a novel contiguous gene syndrome. (J Med Genet. 2000 Feb;37(2):128-31.).

SGPr290.2 SEQ ID NO:11 2p23

Familial syndromic esophageal atresia (Am J Hum Genet. 2000 Feb;66(2):436-44.). Chromosomal rearrangements in acute myelogenous leukemia involving loci on chromosome (Leukemia. 1999 Oct;13(10):1534-8.). Association and linkage analysis of candidate chromosomal regions in multiple sclerosis: indication of disease genes in disease genes in 12q23 and 7ptr-15 (Eur J Hum Genet. 1999 Feb-Mar;7(2):110-6.).

SGPr116 SEQ ID NO:12 6p12

Familial patent ductus arteriosus and bicuspid aortic valve with hand anomalies: a novel heart-hand syndrome. (Am J Med Genet. 1999 Nov 19;87(2):175-9.) Char syndrome, an inherited disorder with patent ductus arteriosus, maps to chromosome 6p12-p21. (Circulation. 1999 Jun 15;99(23):3036-42.). Clinical features of autosomal dominant congenital nystagmus linked to chromosome 6p12. (Am J Ophthalmol. 1998 Jan;125(1):64-70.). Linkage analysis of candidate regions for coeliac disease genes. (Hum Mol Genet. 1997 Aug;6(8):1335-9.). Fine mapping of MEP1A, the gene encoding the alpha subunit of the metalloendopeptidase meprin, to human chromosome 6P21. (Biochem Biophys Res Commun. 1995 Nov 13;216(2):630-5.). Genetic linkage studies in familial frontal epilepsy: exclusion of the human chromosome regions homologous to the El-1 mouse locus. (Epilepsy Res. 1995 Nov;22(3):227-33.)

SGPr003 SEQ ID NO:13 2q37

The expression of fragile sites in lymphocytes of patients with rectum cancer and their first-degree relatives. (Cancer Lett. 2000 May 1;152(2):201-9.). Anterior chamber eye anomalies, redundant skin and syndactyly--a new syndrome associated with breakpoints at 2q37.2 and 7q36.3. (Clin Dysmorphol. 1999 Jul;8(3):157-63.). Wilms' tumor and gonadal dysgenesis in a child with the 2q37.1 deletion syndrome. (Clin Genet. 1998 Apr;53(4):278-80.). A case of Albright's hereditary

osteodystrophy-like syndrome complicated by several endocrinopathies: normal Gs alpha gene and chromosome 2q37. (J Clin Endocrinol Metab. 1998 May;83(5):1563-5.). Albright hereditary osteodystrophy and del(2) (q37.3) in four unrelated individuals. (Am J Med Genet. 1995 Jul 31;58(1):1-7.).

Oguchi disease: suggestion of linkage to markers on chromosome 2q. (J Med Genet. 1995 May;32(5):396-8.). Malformation syndrome with t(2;22) in a cancer family with chromosome instability. (Cancer Genet Cytogenet. 1989 Apr;38(2):223-7.).

SGPr016 SEQ ID NO:14 8p11.1

FGFR1 and MOZ, two key genes involved in malignant hemopathies linked to rearrangements within the chromosomal region 8p11-12. (Bull Cancer. 2000 Dec;87(12):887-94. Review). 5q11, 8p11, and 10q22 are recurrent chromosomal breakpoints in prostate cancer cell lines. (Genes Chromosomes Cancer. 2001 Feb;30(2):187-95.). Unusual breakpoint distribution of 8p abnormalities in T-prolymphocytic leukemia: a study with YACS mapping to 8p11-p12. (Cancer Genet Cytogenet. 2000 Sep;121(2):128-32). Loss of heterozygosity at chromosome segments 8p22 and 8p11.2-21.1 in transitional-cell carcinoma of the urinary bladder. (Int J Cancer. 2000 May 15;86(4):501-5).

SGPr352 SEQ ID NO:15 19p13.3

Clinical characteristics of hereditary cerebrovascular disease in a large family from Colombia (Rev Neurol. 2000 Nov 16-30;31(10):901-7.). Molecular genetic alterations in hamartomatous polyps and carcinomas of patients with Peutz-Jeghers syndrome. (J Clin Pathol. 2001 Feb;54(2):126-31.). Identification of a subtle t(16;19)(p13.3;p13.3) in an infant with multiple congenital abnormalities using a 12-colour multiplex FISH telomere assay, M-TEL. (Eur J Hum Genet. 2000 Dec;8(12):903-10.). Identification of a locus for autosomal dominant polycystic liver disease, on chromosome 19p13.2-13.1. (Am J Hum Genet. 2000 Dec;67(6):1598-604.). Fine mapping of a distinctive autosomal dominant vacuolar

neuromyopathy using 11 novel microsatellite markers from chromosome band 19p13.3. (Eur J Hum Genet. 2000 Oct;8(10):809-12.). Genomewide scan in german families reveals evidence for a novel psoriasis-susceptibility locus on chromosome 19p13. (Am J Hum Genet. 2000 Oct;67(4):1020-4.). Genomewide Search in Canadian Families with Inflammatory Bowel Disease Reveals Two Novel Susceptibility Loci. (Am J Hum Genet. 2000 Jun;66(6):1863-1870.)

SGPr050 SEQ ID NO:16 5q15.3

Mucolipidosis type IV: Novel MCOLN1 mutations in Jewish and non-Jewish patients and the frequency of the disease in the Ashkenazi Jewish population. (Hum Mutat. 2001 May;17(5):397-402.) Myocarditis, a Rare but Severe Manifestation of Q Fever: Report of 8 Cases and Review of the Literature. (Clin Infect Dis. 2001 May 15;32(10):1440-1447.)

SGPr282 SEQ ID NO:17 16p12.3

Linkage of benign familial infantile convulsions to chromosome 16p12-q12 suggests allelism to the infantile convulsions and choreoathetosis syndrome. (Am J Hum Genet. 2001 Mar;68(3):788-94.). A second-generation genomewide screen for asthma-susceptibility alleles in a founder population. (Am J Hum Genet. 2000 Nov;67(5):1154-62.). Evidence of further genetic heterogeneity in autosomal dominant medullary cystic kidney disease. (Nephrol Dial Transplant. 2000 Jun;15(6):818-21.) Localization of a gene for familial juvenile hyperuricemic nephropathy causing underexcretion-type gout to 16p12 by genome-wide linkage analysis of a large family (Arthritis Rheum. 2000 Apr;43(4):925-9.). Localization of a hereditary neuroblastoma predisposition gene to 16p12-p13 (Med Pediatr Oncol. 2000 Dec;35(6):526-30.). Identifying genes predisposing to atopic eczema (J Allergy Clin Immunol. 1999 ov;104(5):1066-70.). Molecular genetics of the neuronal ceroid lipofuscinoses. (Epilepsia. 1999;40 Suppl 3:29-32.). Thirty years of Batten disease research: present status and future goals.

(Mol Genet Metab. 1999 Apr;66(4):231-3.).

SGPr046 SEQ ID NO:18 16q23

A genome-wide family-based linkage study of coeliac disease. (Ann Hum Genet. 2000 Nov;64(Pt 6):479-90.). Pleiotropic syndrome of dehydrated hereditary stomatocytosis, pseudohyperkalemia, and perinatal edema maps to 16q23-q24. (Blood. 2000 Oct 1;96(7):2599-605.). Identification and fine mapping of a region showing a high frequency of allelic imbalance on chromosome 16q23.2 that corresponds to a prostate cancer susceptibility locus. (Cancer Res. 2000 Jul 1;60(13):3645-9.). Concurrent and independent genetic alterations in the stromal and epithelial cells of mammary carcinoma: implications for tumorigenesis. (Cancer Res. 2000 May 1;60(9):2562-6.). A 700-kb physical map of a region of 16q23.2 homozygously deleted in multiple cancers and spanning the common fragile site FRA16D. (Cancer Res. 2000 Mar 15;60(6):1690-7.). Prognostic significance of allelic imbalance of chromosome arms 7q, 8p, 16q, and 18q in stage T3N0M0 prostate cancer. (Genes Chromosomes Cancer. 1998 Feb;21(2):131-43.) Loss of heterozygosity at 16q24.1-q24.2 is significantly associated with metastatic and aggressive behavior of prostate cancer. (Cancer Res. 1997 Aug 15;57(16):3356-9.).

SGPr060 SEQ ID NO:19 15q26

A genome-wide search for susceptibility genes in human systemic lupus erythematosus sib-pair families. (Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14875-9.). Linkage analysis of candidate regions for coeliac disease genes. (Hum Mol Genet. 1997 Aug;6(8):1335-9.).

SGPr068 SEQ ID NO:20 10q22

Autosomal dominant myofibrillar myopathy with arrhythmogenic right ventricular cardiomyopathy linked to chromosome 10q. (Ann Neurol. 1999 Nov;46(5):684-92.) Construction of a high-resolution physical map of the chromosome 10q22-q23 dilated cardiomyopathy locus and analysis of candidate genes. (Genomics. 2000 Jul

15;67(2):109-27.). Chromosomal basis of adenocarcinoma of the prostate. (Cancer Invest. 1999;17(6):441-7.) Allele loss in colorectal cancer at the Cowden disease/juvenile polyposis locus on 10q (Cancer Genet Cytogenet. 1997 Aug;97(1):64-9.) Identification of a genetic locus for familial atrial fibrillation. (N Engl J Med. 1997 Mar 27;336(13):905-11.).

SGPr096 SEQ ID NO:21 3p14

The relationship between genetic susceptibility to head and neck cancer with the expression of common fragile sites. (Head Neck. 2000 Sep;22(6):591-8.). Concurrent and independent genetic alterations in the stromal and epithelial cells of mammary carcinoma: implications for tumorigenesis. (Cancer Res. 2000 May 1;60(9):2562-6.) Prognostic implication of microsatellite alteration profiles in early-stage non-small cell lung cancer. (Clin Cancer Res. 2000 Feb;6(2):559-65). Loss of heterozygosity at chromosomes 3, 6, 8, 11, 16, and 17 in ovarian cancer: correlation to clinicopathological variables. (Cancer Genet Cytogenet. 2000 Oct 1;122(1):49-54.).

SGPr119 SEQ ID NO:22 12q11

Fine genetic mapping of diffuse non-epidermolytic palmoplantar keratoderma to chromosome 12q11-q13: exclusion of the mapped type II keratins. (Exp Dermatol. 1999 Oct;8(5):388-91.).

SGPr143 SEQ ID NO:23 20p13

Hallervorden-Ppatz disease (OMIM 234200).

SGPr164 SEQ ID NO:24 11q25

Deletion mapping of chromosome segment 11q24-q25, exhibiting extensive allelic loss in early onset breast cancer. (Int J Cancer. 2001 Apr 15;92(2):208-13.). Restricted chromosome breakpoint sites on 11q22-q23.1 and 11q25 in various hematological malignancies without MLL/ALL-1 gene rearrangement. (Cancer

Genet Cytogenet. 2001 Jan 1;124(1):27-35.). Autozygosity mapping, to chromosome 11q25, of a rare autosomal recessive syndrome causing histiocytosis, joint contractures, and sensorineural deafness. (Am J Hum Genet. 1998 May;62(5):1123-8.). Tertiary trisomy (22q11q),47,+der(22),t(11;22). (Hum Genet. 1980 Feb;53(2):173-7.).

SGPr281 SEQ ID NO:25 5q31

Interleukin-5 is at 5q31 and is deleted in the 5q- syndrome. (Blood. 1988 Apr;71(4):1150-2.). Lack of association between the interferon regulatory factor-1 (IRF1) locus at 5q31.1 and multiple sclerosis in Germany, northern Italy, Sardinia and Sweden. (Genes Immun. 2000;1(4):290-2.). Childhood asthma: aspects of global environment, genetics and management. (Changgeng Yi Xue Za Zhi. 2000 Nov;23(11):641-61. Review.). Association and linkage of atopic dermatitis with chromosome 13q12-14 and 5q31-33 markers. J Invest Dermatol. 2000 Nov;115(5):906-8. Deletion of 5q31 is observed in megakaryocytic cells in patients with myelodysplastic syndromes and a del(5q), including the 5q- syndrome. (Genes Chromosomes Cancer. 2000 Dec;29(4):350-2.). Ethnic differences in genetic susceptibility to atopy and asthma in Singapore. (Ann Acad Med Singapore. 2000 May;29(3):346-50. Review.). Genomewide scan for prostate cancer-aggressiveness loci. (Am J Hum Genet. 2000 Jul;67(1):92-9.). Molecular genetic analysis of malignant ovarian germ cell tumors. (Gynecol Oncol. 2000 May;77(2):283-8.).

SGPr075 SEQ ID NO:26 Unmapped

SGPr292.2 SEQ ID NO:27 10q26

Sequence homology between 4qter and 10qter loci facilitates the instability of subtelomeric KpnI repeat units implicated in facioscapulohumeral muscular dystrophy. (Am J Hum Genet. 1998 Jul;63(1):181-90.) Frequent loss of heterozygosity on chromosome 10q in muscle-invasive transitional cell carcinomas

of the bladder. (Oncogene. 1997 Jun 26;14(25):3059-66.). Allelic loss on chromosome 10 in prostate adenocarcinoma. (Cancer Res. 1996 May 1;56(9):2143-7.) Severe midline fusion defects in a newborn with 10q26----qter deletion. (Ann Genet. 1989;32(2):124-5.)

SGPr069 SEQ ID NO:28 1p36.3

Neurodevelopmental profile of a new dysmorphic syndrome associated with submicroscopic partial deletion of 1p36.3. (Dev Med Child Neurol. 2000 Mar;42(3):201-6.). Molecular Cytogenetics in Ewing Tumors: Diagnostic and Prognostic Information. (Onkologie. 2000 Oct;23(5):416-422.). Significance of the small subtelomeric area of chromosome 1 (1p36.3) in the progression of malignant melanoma: FISH deletion screening with YAC DNA probes. (Virchows Arch. 1999 Aug;435(2):105-11). Allelic loss on chromosome 1 is associated with tumor progression of cervical carcinoma. (Cancer. 1999 Oct 1;86(7):1294-8). Terminal deletion, del(1)(p36.3), detected through screening for terminal deletions in patients with unclassified malformation syndromes. (Am J Med Genet. 1999 Jan 29;82(3):249-53). Partial monosomy of chromosome 1p36.3: characterization of the critical region and delineation of a syndrome. (Am J Med Genet. 1995 Dec 4;59(4):467-75). Consistent association of 1p loss of heterozygosity with pheochromocytomas from patients with multiple endocrine neoplasia type 2 syndromes. (Cancer Res. 1992 Feb 15;52(4):770-4.).

SGPr212 SEQ ID NO:29 9q22

Chromosome 9 deletions and recurrence of superficial bladder cancer: identification of four regions of prognostic interest. (Oncogene. 2000 Dec 14;19(54):6317-23). Exclusion of NFIL3 as the gene causing hereditary sensory neuropathy type I by mutation analysis. (Hum Genet. 2000 Jun;106(6):594-6). Chromosomal imbalances are associated with a high risk of progression in early invasive (pT1) urinary bladder cancer (Cancer Res. 1999 Nov 15;59(22):5687-91). Brachydactyly type B: linkage to chromosome 9q22 and evidence for genetic heterogeneity. (Am J Hum Genet.

1999 Feb;64(2):578-85). A YAC-based transcript map of human chromosome 9q22.1-q22.3 encompassing the loci for hereditary sensory neuropathy type I and multiple self-healing squamous epithelioma. (Genomics. 1998 Jul 15;51(2):277-81). Molecular analysis of childhood primitive neuroectodermal tumors defines markers associated with poor outcome. (J Clin Oncol. 1998 Jul;16(7):2478-85). Mutilating neuropathic ulcerations in a chromosome 3q13-q22 linked Charcot-Marie-Tooth disease type 2B family. (J Neurol Neurosurg Psychiatry. 1997 Jun;62(6):570-3).

SGPr049 SEQ ID NO:30 5q23.3 / 5q31

Interleukin-5 is at 5q31 and is deleted in the 5q- syndrome. (Blood. 1988 Apr;71(4):1150-2.). Lack of association between the interferon regulatory factor-1 (IRF1) locus at 5q31.1 and multiple sclerosis in Germany, northern Italy, Sardinia and Sweden. (Genes Immun. 2000;1(4):290-2.). Childhood asthma: aspects of global environment, genetics and management. (Changgeng Yi Xue Za Zhi. 2000 Nov;23(11):641-61. Review.). Association and linkage of atopic dermatitis with chromosome 13q12-14 and 5q31-33 markers. J Invest Dermatol. 2000 Nov;115(5):906-8. Deletion of 5q31 is observed in megakaryocytic cells in patients with myelodysplastic syndromes and a del(5q), including the 5q- syndrome. (Genes Chromosomes Cancer. 2000 Dec;29(4):350-2.). Ethnic differences in genetic susceptibility to atopy and asthma in Singapore. (Ann Acad Med Singapore. 2000 May;29(3):346-50. Review.). Genomewide scan for prostate cancer-aggressiveness loci. (Am J Hum Genet. 2000 Jul;67(1):92-9.). Molecular genetic analysis of malignant ovarian germ cell tumors. (Gynecol Oncol. 2000 May;77(2):283-8.).

SGPr026 SEQ ID NO:31 1q31

Genomewide search and genetic localization of a second gene associated with autosomal dominant branchio-oto-renal syndrome: clinical and genetic implications. (Am J Hum Genet. 2000 May;66(5):1715-20.). Jumping translocations involving chromosome 1q in a patient with Crohn disease and acute monocytic leukemia: a

review of the literature on jumping translocations in hematological malignancies and Crohn disease (Cancer Genet Cytogenet. 1999 Mar;109(2):144-9. Review). Molecular analysis of childhood primitive neuroectodermal tumors defines markers associated with poor outcome. (J Clin Oncol. 1998 Jul;16(7):2478-85). Mapping a gene (SRN1) to chromosome 1q25-q31 in idiopathic nephrotic syndrome confirms a distinct entity of autosomal recessive nephrosis. (Hum Mol Genet. 1995 Nov;4(11):2155-8).

SGPr203 SEQ ID NO:32 2q37

The expression of fragile sites in lymphocytes of patients with rectum cancer and their first-degree relatives. (Cancer Lett. 2000 May 1;152(2):201-9.). Anterior chamber eye anomalies, redundant skin and syndactyly--a new syndrome associated with breakpoints at 2q37.2 and 7q36.3. (Clin Dysmorphol. 1999 Jul;8(3):157-63.). Wilms' tumor and gonadal dysgenesis in a child with the 2q37.1 deletion syndrome. (Clin Genet. 1998 Apr;53(4):278-80). Albright hereditary osteodystrophy and del(2)(q37.3) in four unrelated individuals. (Am J Med Genet. 1995 Jul 31;58(1):1-7). Oguchi disease: suggestion of linkage to markers on chromosome 2q. (J Med Genet. 1995 May;32(5):396-8). Malformation syndrome with t(2;22) in a cancer family with chromosome instability. (Cancer Genet Cytogenet. 1989 Apr;38(2):223-7).

SGPr157 SEQ ID NO:33 18q22.3

Psychiatric disorder in a familial 15;18 translocation and sublocalization of myelin basic protein of 18q22.3. (Am J Med Genet. 1996 Apr 9;67(2):154-61.).

SGPr154 SEQ ID NO:34 1q32.1

Oncogene amplification in human gliomas: a molecular cytogenetic analysis. (Oncogene. 1994 Sep;9(9):2717-22).

SGPr088 SEQ ID NO:35 18q23

Molecular characterization of patients with 18q23 deletions. (Am J Hum Genet. 1997 Apr;60(4):860-8.) Unbalanced translocation, t(18;21), detected by fluorescence in situ hybridization (FISH) in a child with 18q- syndrome and a ring chromosome 21. (Am J Med Genet. 1993 Jul 1;46(6):647-51).

EXAMPLE 8b: Candidate Single Nucleotide Polymorphisms (SNPs)

Materials And Methods

The most common variations in human DNA are single nucleotide polymorphisms (SNPs), which occur approximately once every 100 to 300 bases. Because SNPs are expected to facilitate large-scale association genetics studies, there has recently been great interest in SNP discovery and detection. Candidate SNPs for the genes in this patent were identified by blastn searching the nucleic acid sequences against the public database of sequences containing documented SNPs (dbSNP, at NCBI, http://www.ncbi.nlm.nih.gov/SNP/snpblastpretty.html). dbSNP accession numbers for the SNP-containing sequences are given. SNPs were also identified by comparing several databases of expressed genes (dbEST, NRNA) and genomic sequence (i.e., NRNA) for single basepair mismatches. The results are shown in Table 1, in the column labeled "SNPs". These are candidate SNPs – their actual frequency in the human population was not determined. The code below is standard for representing DNA sequence:

G = Guanosine
A = Adenosine
T = Thymidine
C = Cytidine
R = G or A, puRine
Y = C or T, pYrimidine
K = G or T, Keto
W = A or T, Weak (2 H-bonds)

S = C or G, Strong (3 H-bonds)
M = A or C, aMino
B = C, G or T (i.e., not A)
D = A, G or T (i.e., not C)
H = A, C or T (i.e., not G)
V = A, C or G (i.e., not T)
N = A, C, G or T, aNy
X = A, C, G or T

complementary	G	A	T	C	R	Y	W	S	K	M	B	V	D	H	N	X
DNA	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+
strands	C	T	A	G	Y	R	S	W	M	K	V	B	H	D	N	X

For example, if two versions of a gene exist, one with a "C" at a given position, and a second one with a "T: at the same position, then that position is represented as a Y, which means C or T. SNPs may be important in identifying heritable traits associated with a gene.

Results

The results of SNP identification are contained in Table 2 above, and in Example 1, under the section entitled DESCRIPTION OF NOVEL PROTEASE POLYNUCLEOTIDES. As discussed above, a varietyof SNPs were identified in the protease polynucleotides of the present invention.

EXAMPLE 9: Demonstration Of Gene Amplification By Southern Blotting

Materials and Methods

Nylon membranes are purchased from Boehringer Mannheim. Denaturing solution contains 0.4 M NaOH and 0.6 M NaCl. Neutralization solution contains 0.5 M Tris-HCL, pH 7.5 and 1.5 M NaCl. Hybridization solution contains 50% formamide, 6X SSPE, 2.5X Denhardt's solution, 0.2 mg/mL denatured salmon

DNA, 0.1 mg/mL yeast tRNA, and 0.2 % sodium dodecyl sulfate. Restriction enzymes are purchased from Boehringer Mannheim. Radiolabeled probes are prepared using the Prime-it II kit by Stratagene. The β-actin DNA fragment used for a probe template is purchased from Clontech.

Genomic DNA is isolated from a variety of tumor cell lines (such as MCF-7, MDA-MB-231, Calu-6, A549, HCT-15, HT-29, Colo 205, LS-180, DLD-1, HCT-116, PC3, CAPAN-2, MIA-PaCa-2, PANC-1, AsPc-1, BxPC-3, OVCAR-3, SKOV3, SW 626 and PA-1, and from two normal cell lines.

A 10 μg aliquot of each genomic DNA sample is digested with EcoR I restriction enzyme and a separate 10 μg sample is digested with Hind III restriction enzyme. The restriction-digested DNA samples are loaded onto a 0.7% agarose gel and, following electrophoretic separation, the DNA is capillary-transferred to a nylon membrane by standard methods (Sambrook, J. *et al.* (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory).

EXAMPLE 10: Detection Of Protein-Protein Interaction Through Phage Display

Materials And Methods

Phage display provides a method for isolating molecular interactions based on affinity for a desired bait. cDNA fragments cloned as fusions to phage coat proteins are displayed on the surface of the phage. Phage(s) interacting with a bait are enriched by affinity purification and the insert DNA from individual clones is analyzed.

T7 Phage Display Libraries

All libraries were constructed in the T7Select1-1b vector (Novagen) according to the manufacturer's directions.

Bait Presentation

Protein domains to be used as baits are generated as C-terminal fusions to GST and expressed in *E. coli*. Peptides are chemically synthesized and biotinylated at the N-terminus using a long chain spacer biotin reagent.

Selection

Aliquots of refreshed libraries (10^{10}-10^{12} pfu) supplemented with PanMix and a cocktail of *E. coli* inhibitors (Sigma P-8465) are incubated for 1-2 hrs at room temperature with the immobilized baits. Unbound phage is extensively washed (at least 4 times) with wash buffer.

After 3-4 rounds of selection, bound phage is eluted in 100 μL of 1% SDS and plated on agarose plates to obtain single plaques.

Identification of insert DNAs

Individual plaques are picked into 25 μL of 10 mM EDTA and the phage is disrupted by heating at 70 °C for 10 min. 2 μL of the disrupted phage are added to 50 μL PCR reaction mix. The insert DNA is amplified by 35 rounds of thermal cycling (94 °C, 50 sec; 50 °C, 1min; 72 °C, 1min).

Composition of Buffer

10x PanMix

5% Triton X-100

10% non-fat dry milk (Carnation)

10 mM EGTA

250 mM NaF

250 μg/mL Heparin (sigma)

250 μg/mL sheared, boiled salmon sperm DNA (sigma)

0.05% Na azide

Prepared in PBS

Wash Buffer

PBS supplemented with:

0.5% NP-40

25 µl g/mL heparin

PCR reaction mix

1.0 mL 10x PCR buffer (Perkin-Elmer, with 15 mM Mg)

0.2 mL each dNTPs (10 mM stock)

0.1 mL T7UP primer (15 pmol/µL) GGAGCTGTCGTATTCCAGTC

0.1 mL T7DN primer (15 pmol/µL)

AACCCCTCAAGACCCGTTTAG

0.2 mL 25 mM $MgCl_2$ or $MgSO_4$ to compensate for EDTA

Q.S. to 10 mL with distilled water

Add 1 unit of Taq polymerase per 50 µL reaction

LIBRARY: T7 Select1-H441

CONCLUSION

One skilled in the art would readily appreciate that the present invention is well adapted to carry out the objects and obtain the ends and advantages mentioned, as well as those inherent therein. The molecular complexes and the methods, procedures, treatments, molecules, specific compounds described herein are presently representative of preferred embodiments, are exemplary, and are not intended as limitations on the scope of the invention. It will be readily apparent to one skilled in the art that varying substitutions and modifications may be made to the invention disclosed herein without departing from the scope and spirit of the invention.

All patents and publications mentioned in the specification are indicative of the levels of those skilled in the art to which the invention pertains. All patents and publications are herein incorporated by reference to the same extent as if each individual publication was specifically and individually indicated to be incorporated by reference.

The invention illustratively described herein suitably may be practiced in the absence of any element or elements, limitation or limitations which is not specifically disclosed herein. Thus, for example, in each instance herein any of the terms "comprising," "consisting essentially of" and "consisting of" may be replaced with either of the other two terms. The terms and expressions which have been employed are used as terms of description and not of limitation, and there is no intention that in the use of such terms and expressions of excluding any equivalents of the features shown and described or portions thereof, but it is recognized that various modifications are possible within the scope of the invention claimed. Thus, it should be understood that although the present invention has been specifically disclosed by preferred embodiments and optional features, modification and variation of the concepts herein disclosed may be resorted to by those skilled in the art, and that such modifications and variations are considered to be within the scope of this invention as defined by the appended claims.

In addition, where features or aspects of the invention are described in terms of Markush groups, those skilled in the art will recognize that the invention is also thereby described in terms of any individual member or subgroup of members of the Markush group. For example, if X is described as selected from the group consisting of bromine, chlorine, and iodine, claims for X being bromine and claims for X being bromine and chlorine are fully described.

In view of the degeneracy of the genetic code, other combinations of nucleic acids also encode the claimed peptides and proteins of the invention. For example, all four nucleic acid sequences GCT, GCC, GCA, and GCG encode the amino acid alanine. Therefore, if for an amino acid there exists an average of three codons, a polypeptide of 100 amino acids in length will, on average, be encoded by 3100, or 5 x 1047, nucleic acid sequences. Thus, a nucleic acid sequence can be modified to form a second nucleic acid sequence, encoding the same polypeptide as encoded by the first nucleic acid sequences, using routine procedures and without undue experimentation. Thus, all possible nucleic acids that encode the claimed peptides and proteins are also fully described herein, as if all were written out in full taking

into account the codon usage, especially that preferred in humans. Furthermore, changes in the amino acid sequences of polypeptides, or in the corresponding nucleic acid sequence encoding such polypeptide, may be designed or selected to take place in an area of the sequence where the significant activity of the polypeptide remains unchanged. For example, an amino acid change may take place within a β-turn, away from the active site of the polypeptide. Also changes such as deletions (*e.g.* removal of a segment of the polypeptide, or in the corresponding nucleic acid sequence encoding such polypeptide, which does not affect the active site) and additions (*e.g.* addition of more amino acids to the polypeptide sequence without affecting the function of the active site, such as the formation of GST-fusion proteins, or additions in the corresponding nucleic acid sequence encoding such polypeptide without affecting the function of the active site) are also within the scope of the present invention. Such changes to the polypeptides can be performed by those with ordinary skill in the art using routine procedures and without undue experimentation. Thus, all possible nucleic and/or amino acid sequences that can readily be determined not to affect a significant activity of the peptide or protein of the invention are also fully described herein.

The invention has been described broadly and generically herein. Each of the narrower species and subgeneric groupings falling within the generic disclosure also form part of the invention. This includes the generic description of the invention with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

Other embodiments are within the following claims.

What is claimed is:

CLAIMS

1. An isolated, enriched or purified nucleic acid molecule encoding a protease polypeptide, wherein said nucleic acid molecule comprises a nucleotide sequence that:

(a) encodes a polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70;

(b) is the complement of the nucleotide sequence of (a);

(c) hybridizes under stringent conditions to the nucleotide molecule of (a) and encodes a protease polypeptide.

2. The nucleic acid molecule of claim 1, further comprising a vector or promoter effective to initiate transcription in a host cell.

3. The nucleic acid molecule of claim 1, wherein said nucleic acid molecule is isolated, enriched, or purified from a mammal.

4. The nucleic acid molecule of claim 3, wherein said mammal is a human.

5. A nucleic acid molecule of claim 1 comprising a nucleic acid having a nucleotide sequence which hybridizes under stringent conditions to a nucleotide

sequence encoding a protease polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

6. An isolated, enriched, or purified protease polypeptide, wherein said polypeptide comprises an amino acid sequence at least about 90% identical to a sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

7. The protease polypeptide of claim 7, wherein said polypeptide is isolated, purified, or enriched from a mammal.

8. The protease polypeptide of claim 8, wherein said mammal is a human.

9. An antibody or antibody fragment having specific binding affinity to a protease polypeptide or to a domain of said polypeptide, wherein said polypeptide comprises an amino acid sequence selected from the group consisting of those set

forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

10. A hybridoma which produces the antibody of Claim 9.

11. A kit comprising an antibody which binds to a polypeptide of claim 6 and a negative control antibody.

12. A method for identifying a substance that modulates the activity of a protease polypeptide comprising the steps of:

(a) contacting the protease polypeptide substantially identical to an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70 with a test substance;

(b) measuring the activity of said polypeptide; and

(c) determining whether said substance modulates the activity of said polypeptide.

13. A method for identifying a substance that modulates the activity of a protease polypeptide in a cell comprising the steps of:

(a) expressing a protease polypeptide having substantially identical to an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70;

(b) adding a test substance to said cell; and

(c) monitoring a change in cell phenotype or the interaction between said polypeptide and a natural binding partner.

14. A method for treating a disease or disorder by administering to a patient in need of such treatment a substance that modulates the activity of a protease substantially identical to an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

15. The method of claim 15, wherein said disease or disorder is selected from the group consisting of cancers, immune-related diseases and disorders,

cardiovascular disease, brain or neuronal-associated diseases, metabolic disorders and inflammatory disorders.

16. The method of claim 15, wherein said disease or disorder is selected from the group consisting of cancers of tissues; cancers of blood or hematopoietic origin; cancers of the breast, colon, lung, prostrate, cervical, brain, ovarian, bladder or kidney.

17. The method of claim 15, wherein said disease or disorder is selected from the group consisting of central or peripheral nervous system diseases, migraines; pain; sexual dysfunction; mood disorders; attention disorders; cognition disorders; hypotension; hypertension; psychotic disorders; neurological disordersand dyskinesias.

18. The method of claim 15, wherein said substance modulates protease activity *in vitro*.

19. The method of claim 19, wherein said substance is a protease inhibitor.

20. A method for detection of a protease polypeptide in a sample as a diagnostic tool for a disease or disorder, wherein said method comprises:

(a) contacting said sample with a nucleic acid probe which hybridizes under hybridization assay conditions to a nucleic acid target region of a protease polypeptide having an amino acid sequence selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59,

SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70, or one or more fragments thereof, with a control nucleic acid target region encoding said protease polypeptide, or one or more fragments thereof; and

(b) detecting differences in sequence or amount between said target region and said control target region, as an indication of said disease or disorder.

25. The method of claim 25, wherein said disease or disorder is selected from the group consisting of cancers, immune-related diseases and disorders, cardiovascular disease, brain or neuronal-associated diseases, metabolic disorders and inflammatory disorders.

26. The method of claim 25, wherein said disease or disorder is selected from the group consisting of cancers of tissues; cancers of hematopoietic cancers of blood or hematopoietic origin; cancers of the breast, colon, lung, prostrate, cervical, brain, ovarian, bladder or kidney.

27. The method of claim 25, wherein said disease or disorder is selected from the group consisting of central or peripheral nervous systems disease, migraines, pain; sexual dysfunction; mood disorders; attention disorders; cognition disorders; hypotension; hypertension; psychotic disorders; neurological disorders; and dyskinesias.

28. An isolated, enriched or purified nucleic acid molecule that comprises a nucleic molecule encoding a domain of a protease polypeptide having a sequence selected from the group consisting of SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52,

SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

29. An isolated, enriched or purified nucleic acid molecule encoding a protease polypeptide which comprises a nucleotide sequence that encodes a polypeptide having an amino acid sequence that has least 90 % identity to a polypeptide selected from the group consisting of those set forth in SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

30. The isolated, enriched or purified nucleic acid molecule according to Claim 1 wherein the molecule comprises a nucleotide sequence substantially identical to a sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35.

31. An isolated, enriched or purified nucleic acid molecule consisting essentially of about 10-30 contiguous nucleotide bases of a nucleic acid sequence

that encodes a polypeptide that is selected from the group consisting of SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, SEQ ID NO:39, SEQ ID NO:40, SEQ ID NO:41, SEQ ID NO:42, SEQ ID NO:43, SEQ ID NO:44, SEQ ID NO:45, SEQ ID NO:46, SEQ ID NO:47, SEQ ID NO:48, SEQ ID NO:49, SEQ ID NO:50, SEQ ID NO:51, SEQ ID NO:52, SEQ ID NO:53, SEQ ID NO:54, SEQ ID NO:55, SEQ ID NO:56, SEQ ID NO:57, SEQ ID NO:58, SEQ ID NO:59, SEQ ID NO:60, SEQ ID NO:61, SEQ ID NO:62, SEQ ID NO:63, SEQ ID NO:64, SEQ ID NO:65, SEQ ID NO:66, SEQ ID NO:67, SEQ ID NO:68, SEQ ID NO:69, and SEQ ID NO:70.

32. The isolated, enriched or purified nucleic acid molecule of Claim 31 consisting essentially of about 10-30 contiguous nucleotide bases of a nucleic acid sequence selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, SEQ ID NO:3, SEQ ID NO:4, SEQ ID NO:5, SEQ ID NO:6, SEQ ID NO:7, SEQ ID NO:8, SEQ ID NO:9, SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, SEQ ID NO:18, SEQ ID NO:19, SEQ ID NO:20, SEQ ID NO:21, SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, SEQ ID NO:27, SEQ ID NO:28, SEQ ID NO:29, SEQ ID NO:30, SEQ ID NO:31, SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35.

Figure 1A

>SGPR_140_SEQ ID NO:1
ATGAGGGGCCTTGTGGTATTCCTTGCAGTCTTTGCTCTCTCTGAGGTCAATGCCATCACCAGGGTTC
CTCTGCACAAAGGGAAGTCGCTGAGGAGGGCCCTGAAGGAGCGCAGGCTCCTGGAGGACTTCCTG
AGGAATCACCATTATGCAGTCAGCAGGAAGCACTCCAGCTCTGGGGTGGTGGCCAGCGAGTCTCT
GACCAACTACCTGGATTGTCAGTACTTTGGGAAGATCTACATCGGGACCCTTCCCCAGAAGTTCAC
CTTGGTGTTTGATACAGGCTCCCCGGATATCTGGGTGCCCTCTGTCTACTGCAACAGTGATGCCTGT
CAGAACCACCAACGCTTCGATCCGTCCAAGTCCTCCACCCAGAACATGGGCAAGTCCCTGTCCATC
CAGTATGGCACAGGCAGCATGCGGGGCTTGCTGGGCTATGACACTGTCACCGTCTCCAACATTGTG
GACCCCCACCAGACTGTGGGTCTGAGCACCCAGGAACCTGGCGACGTCTTCACCTACTCCGAGTTT
GATGGGATCCTGGGGCTGGCCTATCCCTCTCTTGCCTCTGAGTACGCGCTGCGCCTTGGTTTCAGG
AATGACCAGGGGAGCATGCTCACGCTGAGGGCCATTGATCTGTCGTACTACACAGGCTCCCTGCAC
TGGATACCCATGACTGCAAGAATACTGGCAGTTCACTGTGGACAGGAAGGACCTGGGGAGGGAGG
GCTGGATGAGGCCATCTTGCATACCTTTGGAAGTGTCATCATTGACGGCGTGGTGGTGGCCTGTGA
CGGTGGCTGTCAGGCCATCCTGGACACCGGCACCTCCCTGCTGGTGGGGCCTGGTGGCAACATCCT
CAACATCCAGCAGGCCATTGGACGCACTGCGGGCCAGTACAATGAGTTTGACATCGACTGCGGGC
GCCTGAGCAGCATTCCCACGGCTGTCTTCGAGATCCACGGCAAGAAGTACCCCCTGCCACCCTCCG
CCTATACCAGCCAGGACCAGGGCTTCTGCACCAGTGGTTTCCAGGGTGACTATAGTTCCCAGCAGT
GGATCCTGGGGAATGTCTTCATCTGGGAGTATTACAGTGTCTTTGACAGGACCAATAACCGTGTGG
GGCTGGCGAAGGCTGTCTGA

>SGPR_197_SEQ ID NO:2
ATGGATAGATGCAAACATGTAGGGCGGTTACGGCTCGCCCAGGACCACTCCATCCTGAACCCTCA
GAAGTGGTGCTGCTTAGAGTGTGCCACCACCGAGTCCGTGTGGGCCTGCCTCAAGTGCTCCCACGT
GGCCTGCGGCCGCTATATTGAGGACCACGCCCTGAAACACTTTGAGGAGACGGGACACCCGCTAG
CCATGGAAGTCCGGGATCTCTACGTGTTCTGTTACCTGTGCAAGGACTACGTGCTCAATGATAACC
CAGAGGGGGACCTGAAGCTGCTAAGAAGCTCCCTCCTGGCGGTCCGGGGCCAGAAACAGGACACG
CCGGTGAGACGTGGGCGGACGCTGCGGTCCATGGCTTCGGGTGAGGACGTGGTCCTGCCGCAGCG
CGCTCCTCAGGGACAGCCGCAGATGCTCACGGCTCTGTGGTACCGGCGTCAGCGCCTGCTGGCCAG
GACGCTGCGGCTGTGGTTCGAGAAGAGCTCCCGGGGCCAGGCGAAGCTGGAGCAGCGGCGGCAG
GAGGAGGCCCTGGAGCGCAAGAAGGAGGAGGCGCGGAGGCGGCGGCGAGCCGGCCATGGCCC
CAGGCGTCACGGGCCTGCGCAACCTGGGCAACACCTGCTACATGAACTCCATCCTCCAGGTGCTCA
GCCACCTCCAGAAGTTCCGAGAATGTTTCCTCAACCTTGACCCTTCCAAAACGGAACATCTGTTTC
CCAAAGCCACCAACGGGAAGACTCAGCTTTCTGGCAAGCCAACCAACAGCTCGGCCACGGAGCTG
TCCTTGAGAAATGACAGGGCCGAGGCATGCGAGCGGGAGGGTTTCTGCTGGAACGGCAGGGCCTC
CATTAGTCGGAGTCTGGAGCTCATCCAGAACAAGGAGCCGAGTTCAAAGCACATTTCCCTCTGCCG
TGAACTGCACACCCTCTTCCGAGTCATGTGGTCCGGGAAGTGGGCCCTAGTGTCGCCCTTCGCCAT
GCTCCACTCAGTGTGGAGCCTGATCCCTGCCTTCCGCGGCTACGACCAACAGGACGCGCAGGAATT
TCTCTGCGAGCTGCTGCACAAGGTGCAGCAGGAACTCGAGTCTGAGGGCACCACACGCCGGATCC
TCATCCCCTTCTCCCAGAGGAAGCTCACCAAACAGGTCTTAAAGGTGGTGAATACCATATTTCATG
GGCAGCTGCTCAGTCAGGGAAGGTGGTCTGGCCGTAATCATCGAGAGAAGATTGGGGTCCATGTC
GTCTTTGACCAGGTATTAACCATGGAACCTTACTGCTGCAGGGACATGCTCTCCTCTCTTGACAAA
GAGACCTTTGCCTATGATCTCTCCGCAGTGGTCATGCATCACGGGAAAGGGTTTGGCTCAGGACAC
TACACAGCCTATTGCTACAACACAGAGGGAGGGGAGCAGACCCAGGGTTTGGCCATCACCAACCG
GGAGTACGGCCTAAGCCAGAGGGAGCTGGCACCACCTTCGAAAGCATTCCCTTTGATGTGA

>SGPR_005_SEQ ID NO:3
ATGGGGCCAAGACTCATTCCGTTTCTATTTTTGTTTGTTTACCCTATTCTCTGCAGGATCATTCTGA
GGAAAGGCAAGTCTATCCGCCAGAGAATGGAGGAGCAGGGTGTACTGGAGACGTTTCTGAGGGAC
CACCCAAAGGCTGATCCAATTGCCAAGTATTATTTCAATAATGATGCTGTTGCTTATGAGCCCTTC
ACCAACTACCTGGATTCTTTCTACTTTGGGGAGATCAGCACTGGGACACCACCCCAAAATTTCCTA
GTCTCTTTGATACGGGTTCCTCCAATCTGTAGCCTGCCCTCCATCTACTGCCAGAGCCAAGTCTGCT
CCAATCACAACAGGTTCAATCCCAGCCTGTCCTCCACCTTCAGAAACGATGGACAAACCTATGGAC
TATCCTATGGGAGTGGCAGCCTGAGTGTGTTCCTGGGCTATGACACTGTGACTGTTCATAACATCG
TTGTCAATAACCAGGAGTTTGGCCTGAGTGAGAATGAGCCCAGCGACCCCTTTTACTATTCAGACT
TTGACGGGATCCTGGGAATGGCCTACCCAAACATGGCAGAGGGGAATTCCCCTACAGTAATGCAG
GGGATGCTGCAGCAGAGCCAGCTTACTCAGCCCGTCTTCAGCTTCTACTTCACCTGCCAGCCAACC
CGCCAGTATTGTGGAGAGCTCATCCTTGGAGGTGTGGACCCCAACCTTTATTCTGGTCAGATCATC
TGGACCCCTGTCAGCCCGGAACTGTACTGGCAGATTGCCATCGAGGAATTTGCCATCGGTAACCAG

Figure 1B

GCCACTGGCTTGTGCTCTGAGGGTTGCCAGGCCATTGTGGATACCGAGACCTTCCTGCTGGCAGTT
CCTCAGCAGTACATGGCCTCCTTCCTGCAGGCAACAGGACCCCAGCAGGCTCAGAATGGTGACTTT
GTGGTCAACTGCAGCTACATACAGAGCATGCCCACCATCACCTTCATCATCGGCGGGGCCCAGTTT
CCTCTGCCTCCCTCTGAATATGTCTTCAATAACAATGGCTACTGCAGGCTTGGAACTGAGGCCACC
TGCCTGCCCTCCCGCAGTGGGCAGCCCCTCTGGATTCTGGGGGATGTCTTCCTCAAGGAATATTGC
TCTGTCTATGACATGGCCAACAACAGGGTGGGCTTTGCCTTCTCTGCCTAG

>SGPR_078_SEQ ID NO:4
ATGCAGCCCTCCAGCCTTCTGCCGCTCGCCCTCTGCCTGCTGGCTGCACCCGCCTCCGCGCTCGTCA
GGATCCCGCTGCACAAGTTCACGTCCATCCGCCGGACCATGTCGGAGGTTGGGGGCTCTGTGGAG
GACCTGATTGCCAAAGGCCCCGTCTCAAAGTACTCCCAGGCGGTGCCAGCCGTGACCGAGGGGCC
CATTCCCGAGGTGCTCAAGAACTACATGGACGCCCAGTACTACGGGGAGATTGGCATCGGGACGC
CCCCCCAGTGCTTCACAGTCGTCTTCGACACGGGCTCCTCCAACCTGTGGGTCCCCTCCATCCACTG
CAAACTGCTGGACATCGCTTGCTGGATCCACCACAAGTACAACAGCGACAAGTCCAGCACCTACG
TTAAGAATGGTACCTCGTTTGACATCCACTATGGCTCGGGCAGCCTCTCCGGGTACCTGAGCCAGG
ACACTGTGTCGGTGCCCTGCCAGTCAGCGTCGTCAGCCTCTGCCCTGGGCGGTGTCAAAGTGGAGA
GGCAGGTCTTTGGGGAGGCCACCAAGCAGCCAGGCATCACCTTCATCGCAGCCAAGTTCGATGGC
ATCCTGGGCATGGCCTACCCCCGCATCTCCGTCAACAACGTGCTGCCCGTCTTCGACAACCTGATG
CAGCAGAAGCTGGTGGACCAGAACATCTTCTCCTTCTACCTGAGCAGGGACCCAGATGCGCAGCC
TGGGGGTGAGCTGATGCTGGGTGGCACAGACTCCAAGTATTACAAGGGTTCTCTGTCCTACCTGAA
TGTCACCCGCAAGGCCTACTGGCAGGTCCACCTGGACCAGGTGGAGGTGGCCAGCGGGCTGACCC
TGTGCAAGGAGGGCTGTGAGGCCATTGTGGACACAGGCACTTCCCTCATGGTGGGCCCGGTGGAT
GAGGTGCGCGAGCTGCAGAAGGCCATCGGGGCCGTGCCGCTGATTCAGGGCGAGTACATGATCCC
CTGTGAGAAGGTGTCCACCCTGCCCGCGATCACACTGAAGCTGGGAGGCAAAGGCTACAAGCTGT
CCCCAGAGGACTACACGCTCAAGGTGTCGCAGGCCGGGAAGACCCTCTGCCTGAGCGGCTTCATG
GGCATGGACATCCCGCCACCCAGCGGGCCACTCTGGATCCTGGGCGACGTCTTCATCGGCCGCTAC
TACACTGTGTTTGACCGTGACAACAACAGGGTGGGCTTCGCCGAGGCTGCCCGCCTCTAG

>SGPR_084_SEQ ID NO:5
ATGGCTCTCCTGACCAATCTACTGCCCCTGTGCTGCTTGGCACTTCTGGCGCTGCCAGCCCAGAGCT
GCGGGCCGGGCCGGGGGCCGGTTGGCCGGCGCCGCTATGCGCGCAAGCAGCTCGTGCCGCTACTC
TACAAGCAATTTGTGCCCGGCGTGCCAGAGCGGACCCTGGGCGCCAGTGGGCCAGCGGAGGGGAG
GGTGGCAAGGGGCTCCGAGCGCTTCCGGGACCTCGTGCCCAACTACAACCCCGACATCATCTTCAA
GGATGAGGAGAACAGTGGAGCCGACCGCCTGATGACCGAGCGTTGTAAGGAGCGGGTGAACGCTT
TGGCCATTGCCGTGATGAACATGTGGCCCGGAGTGCGCCTACGAGTGACTGAGGGCTGGGACGAG
GACGGCCACCACGCTCAGGATTCACTCCACTACGAAGGCCGTGCTTTGGACATCACTACGTCTGAC
CGCGACCGCAACAAGTATGGGTTGCTGGCGCGCCTCGCAGTGGAAGCCGGCTTCGACTGGGTCTA
CTACGAGTCCCGCAACCACGTCCACGTGTCGGTCAAAGCTGATAACTCACTGGCGGTCCGGGCGG
GCGGCTGCTTTCCGGGAAATGCAACTGTGCGCCTGTGGAGCGGCGAGCGGAAAGGGCTGCGGGAA
CTGCACCGCGGAGACTGGGTTTTGGCGGCCGATGCGTCAGGCCGGGTGGTGCCCACGCCGGTGCT
GCTCTTCCTGGACCGGGACTTGCAGCGCCGGGCTTCATTTGTGGCTGTGGAGACCGAGTGGCCTCC
ACGCAAACTGTTGCTCACGCCCTGGCACCTGGTGTTTGCCGCTCGAGGGCCGGCGCCCGCGCCAGG
CGACTTTGCACCGGTGTTCGCGCGCCGGCTACGCGCTGGGGACTCGGTGCTGGCGCCCGGCGGGG
ATGCGCTTCGGCCAGCGCGCGTGGCCCGTGTGGCGCGGGAGGAAGCCGTGGGCGTGTTCGCGCCG
CTCACCGCGCACGGGACGCTGCTGGTGAACGATGTCCTGGCCTCTTGCTACGCGGTTCTGGAGAGT
CACCAGTGGGCGCACCGCGCTTTTGCCCCCTTGAGACTGCTGCACGCGCTAGGGGCGCTGCTCCCC
GGCGGGGCCGTCCAGCCGACTGGCATGCATTGGTACTCTCGGCTCCTCTACCGCTTAGCGGAGGAG
CTACTGGGCTGA

>SGPR_009_SEQ ID NO:6
ATGGCTGAGAAACCATCCAACGGTGTTCTGGTCCACATGGTGAAGTTGCTGATCAAGACCTTTCTA
GATGGCATTTTTGATGATTTGATGGAAAATAATGTATTAAATACAGATGAGATACACCTTATAGGA
AAATGTCTAAAGTTTGTGGTGAGCAATGCTGAAAACCTGGTTGATGATATCACTGAGACAGCTCAA
ACTGCAGGCAAAATATTTAGGGAACACCTGTGGAATTCCAAAAAACAGCTGAGTTCAATTTTTTTC
TCTCTTTCAGCTTTTCTGGAAATCCAGGGTGCCCAACCCAGTGGCAAGTTAAAGCTTTGTCCTCATG
CTCACTTCCATGAACTAAAGACAAAAAGGGCAGATGAGATATATCCAGTGATGGAGAAAGAGAG
GCGAACATGCCTGGGCCTCAACATCCGCAACAAAGAATTCAACTATCTTCATAATCGAAATGGTTC
TGAACTTGACCTTTTGGGGATGCGAGATCTACTTGAAAACCTTGGATACTCAGTGGTTATAAAAGA

Figure 1C

GAATCTCACAGCTCAGGAAATGGAAACAGCACTAAGGCAGTTTGCTGCTCACCCAGAGCACCAGT
CCTCAGACAGCACATTCCTGGTGTTTATGTCACATAGCATCCTGAATGGAATCTGTGGGACCAAGC
ACTGGGATCAAGAGCCAGATGTTCTTCACGATGACACCATCTTTGAAATTTTCAACAACCGTAACT
GCCAGAGTCTGAAAGACAAACCCAAGGTCATCATCATGCAAGCCTGCCGAGGCAATGGTGCTGGG
ATTGTTTGGTTCACCACTGACAGTGGAAAAGCCGGTGCAGATACTCATGGTCGGCTCTTGCAAGGT
AACATCTGTAATGATGCTGTTACAAAGGCTCATGTGGAAAAGGACTTCATTGCTTTCAAATCTTCC
ACACCACATAATGTTTCTTGGAGACATGAAACAAATGGCTCTGTCTTCATTTCCCAAATTATCTACT
ACTTCAGAGAGTATTCTTGGAGTCATCATCTAGAGGAAATTTTTCAAAAGGTTCAACATTCATTTG
AGACCCCAAATATACTGACCCAGCTGCCCACCATTGAAAGACTATCCATGACACGATATTTCTATC
TCTTTCCTGGGAATTAA

>SGPR_286_SEQ ID NO:7
CAGTATGACCTGTCCAAGGCCAGGGCTGCCCTCCTCCTGGCTGTGATCCAAGGCCGGCCTGGGGCC
CAGCATGACGTGGAGGCGCTGGGGGGCCTGTGCTGGGCCCTGGGCTTTGAGACCACCGTGAGAAC
GGACCCTACAGCCCAGGCTTTCCAGGAGGAGCTGGCCCAGTTCCGGGAGCAACTGGACACCTGCA
GGGGCCCTGTGAGCTGTGCCCTTGTGGCCCTGATGGCCCATGGGGGACCACGGGGTCAGCTGCTG
GGGGCTGACGGGCAAGAGGTGCAGCCCGAGGCACTCATGCAGGAGCTGAGCCGCTGCCAGGTGCT
GCAGGGCCGCCCCAAGATCTTCCTGTTGCAGGCCTGCCGTGGGGGAAACAGGGATGCTGGTGTGG
GGCCCACAGCTCTCCCCTGGTACTGGAGCTGGCTGCGGGCACCTCCATCTGTCCCCTCCCATGCAG
ATGTCCTGCAGATCTACGCTGAGGCCCAAGGCTATGTGGCCTATCGCGATGACAAGGGCTCAGACT
TTATCCAGACACTGGTGGAGGTCCTCAGAGCCAACCCCGGGAGAGACCTTCTGGAGCTGCTGACT
GAGGTCAACAGGCGGGTGTGCGAGCAGGAGGTGCTGGGCCCCGACTGCGATGAACTCCGCAAGGC
CTGCCTGGAGATCCGCAGCTCGCTCCGGCGCCGGCTCTGCCTCCAGGCCTGA

>SGPR_008_SEQ ID NO:8
ATGGCGTATTACCAGGAGCCTTCAGTGGAGACCTCCATCATCAAGTTCAAAGACCAGGACTTTACC
ACCTTGCGGGATCACTGCCTGAGCATGGGCCGGACGTTTAAGGATGAGACATTCCCTGCAGCAGA
TTCTTCCATAGGCCAGAAGCTGCTCCAGGAAAAACGCCTCTCCAATGTGATATGGAAGCGGCCAC
AGGATCTACCAGGGGGTCCTCCTCACTTCATCCTGGATGATATAAGCAGATTTGACATCCAACAAG
GAGGCGCAGCTGACTGCTGGTTCCTGGCAGCACTGGGATCCTTGACTCAGAACCCACAGTACAGG
CAGAAGATCCTGATGGTCCAAAGCTTTTCACACCAGTATGCTGGCATTTTCCGTTTCCGGTTCTGGC
AATGTGGCCAGTGGGTGGAAGTGGTGATTGATGACCGCCTACCTGTCCAGGGAGATAAATGCCTC
TTTGTGCGTCCTCGCCACCAAAACCAAGAGTTCTGGCCCTGCCTGCTGGAGAAGGCCTATGCCAAG
CTGCTCGGATCCTATTCCGATCTGCACTATGGCTTCCTCGAGGATGCCCTGGTGGACCTCACAGGA
GGCGTGATCACCAACATCCATCTGCACTCTTCCCCTGTGGACCTGGTGAAGGCAGTGAAGACAGCG
ACCAAGGCAGGCTCCCTGATAACCTGTGCCACTCCAAGTGGGCCAACAGATACAGCACAGGCGAT
GGAGAATGGGCTGGTGAGTCTCCATGCCTACACTGTGACTGGGGCTGAGCAGATTCAATACCGAA
GGGGCTGGGAAGAAATTATCTCCCTGTGGAACCCCTGGGGCTGGGGCGAGGCCGAATGGAGAGGG
CGCTGGAGTGATGGGTCTCAGGAGTGGGAGGAAACCTGTGATCCGCGGAAAAGCCAGCTACATAA
GAAACGGGAAGATGGCGAGTTTTGGATGTCGTGTCAAGATTTCCAACAGAAATTCATCGCCATGTT
TATATGTAGCGAAATTCCAATTACCCTGGACCATGGAAACACACTCCACGAAGGATGGTCCCAAA
TAATGTTTAGGAAGCAAGTGATTCTAGGAAACACTGCAGGAGGACCTCGGAATGATGCTCAATTC
AACTTCTCTGTGCAAGAGCCAATGGAAGGCACCAATGTTGTCGTGTGCGTCACAGTTGCTGTCACA
CCATCAAATTTGAAAGCAGAAGATGCAAAATTTCCACTCGATTTCCAAGTGATTCTGGCTGGCTCA
CAGCGGTTCCGGGAGAAATTTCCACCCGTGTTTTTTTCCTCGTTCAGAAACACTGTCCAAAGCTCA
AATAATAAATTCCGCCGCAACTTCACCATGACTTACCATCTGAGCCCTGGGAACTATGTTGTGGTT
GCACAGACACGGAGAAAATCAGCGGAGTTCTTGCTCCGAATCTTCCTGAAAATGCCAGACAGTGA
CAGGCACCTGAGCAGCCATTTCAACCTCAGAATGAAGGGAAGCCCTTCAGAACATGGCTCCCAAC
AAAGCATTTTCAACAGATATGCTCAGCAGAGGCTGGACATTGATGCCACCCAGCTTCAGGGCCTTC
TCAACCAGGAGCTTCTAACAGGACCTCCAGGGGACATGTTCTCCTTAGATGAGTGCCGCAGCTTGG
TGGCTCTGATGGAACTGAAAGTGAATGGGCGGCTAGACCAAGAGGAGTTTGCGCGACTGTGGAAG
CGCCTTGTTCACTACCAGCATGTTTTCCAGAAGGTTCAGACAAGCCCTGGAGTCCTCCTGAGCTCG
GACTTGTGGAAGGCCATAGAGAATACAGACTTCCTCAGAGGGATCTTCATCAGCCGTGAGCTGCT
GCATCTGGTGACCCTCAGGTACAGCGACAGCGTCGGCAGGGTCAGCTTCCCCAGCCTGGTCTGCTT
CCTGATGCGGCTTGAAGCCATGGCAAAGACCTTCCGCAACCTCTCTAAGGATGGAAAAGGACTCT
ACCTGACAGAAATGGAGTGGATGAGCCTGGTCATGTACAACTGA

Figure 1D

>SGPR_198_SEQ ID NO:9
ATGGCAGCCCAGGCAGCTGGTGTATCTAGGCAGCGGGCAGCCACTCAAGGTCTTGGCTCCAACCA
AAACGCTTTGAAGTACTTGGGCCAGGATTTCAAGACCCTGAGGCAACAGTGCTTGGACTCAGGGG
TCCTATTTAAGGACCCTGAGTTCCCAGCATGTCCATCAGCTTTGGGCTACAAGGATCTTGGACCAG
GCTCTCCGCAAACTCAAGGCATCATCTGGAAGCGGCCCACGGAGTTGTGTCCCAGCCCTCAGTTTA
TCGTTGGTGGAGCCACGCGCACAGACATTTGTCAGGGTGGTCTAGGTGACTGCTGGCTTCTGGCTG
CCATTGCCTCCCTGACCCTGAATGAAGAGCTGCTTTACCGGGTGGTCCCCAGGGACCAGGACTTCC
AGGAGAACTATGCGGGAATCTTTCACTTTCAGTTCTGGCAGTACGGAGAGTGGGTGGAGGTGGTC
ATTGACGACAGGCTGCCCACCAAGAATGGACAGCTGCTCTTCCTACACTCGGAACAAGGCAATGA
ATTCTGGAGTGCCCTGCTGGAGAAAGCCTATGCCAAGCTTAATGGTTGTTATGAGGCTCTCGCTGG
AGGTTCCACAGTGGAGGGGTTTGAGGATTTCACAGGTGGCATCTCTGAGTTTTATGACCTGAAGAA
ACCACCAGCCAATCTATATCAGATCATCCGGAAGGCCCTCTGTGCGGGGTCTCTGCTGGGCTGCTC
CATTGATGTCTACAGTGCAGCCGAAGCCGAAGCCATCACCAGCCAGAAGCTGGTTAAGAGTCATG
CGTACTCTGTCACTGGAGTCGAAGAGGTGAATTTCCAGGGCCATCCAGAGAAGCTGATCAGACTC
AGGAATCCATGGGGTGAAGTGGAGTGGTCGGGAGCCTGGAGCGATGATGCACCAGAGTGGAATC
ACATAGACCCCCGGCGGAAGGAAGAACTGGACAAGAAAGTTGAGGATGGAGAATTCTGGATGTC
ACTTTCAGATTTCGTGAGGCAGTTCTCTCGGTTGGAGATCTGCAACCTGTCCCCGGACTCTCTGAGT
AGCGAGGAGGTGCACAAATGGAACCTGGTCCTGTTCAACGGCCACTGGACCCGGGGCTCCACAGC
TGGGGGCTGCCAGAACTACCCAGCCACGTACTGGACCAATCCCCAGTTCAAAATCCGTTTGGATGA
AGTGGATGAGGACCAGGAGGAGAGCATCGGTGAACCCTGCTGTACAGTGCTGCTGGGCCTGATGC
AGAAAAATCGCAGGTGGCGGAAGCGGATAGGACAAGGCATGCTTAGCATCGGCTATGCCGTCTAC
CAGGTTCCCAAGGAGCTGGAGAGTCACACGGACGCACACTTGGGCCGGGATTTCTTCCTGGCCTAC
CAGCCCTCAGCCCGCACCAGCACCTACGTCAACCTGCGGGAGGTCTCTGGCCGGGCCCGGCTGCCC
CCTGGGGAGTACCTGGTGGTGCCATCCACATTTGAACCCTTCAAAGACGGCGAGTTCTGCTTGAGA
GTGTTCTCAGAGAAGAAGGCCCAGGCCCTAGAAATTGGGGATGTGGTAGCTGGAAACCCATATGA
GCCACATCCCAGTGAGGTGGATCAGGAAGATGACCAGTTCAGGAGGCTGTTTGAGAAGTTGGCAG
GGAAGGATTCTGAGATTACTGCCAATGCACTCAAGATACTTTTGAATGAGGCGTTTTCCAAGAGAA
CAGACATAAAATTCGATGGATTCAACATCAACACTTGCAGGGAAATGATCAGTCTGTTGGATAGC
AATGGAACGGGCACTTTGGGGGCGGTGGAATTCAAGACGCTCTGGCTGAAGATTCAGAAGTATCT
GGAGATCTATTGGGAAACTGATTATAACCACTCGGGCACCATCGATGCCCACGAGATGAGGACAG
CCCTCAGGAAGGCAGGTTTCACCCTCAACAGCCAGGTGCAGCAGACCATTGCCCTGCGGTATGCGT
GCAGCAAGCTCGGCATCAACTTTGACAGCTTCGTGGCTTGTATGATCCGCCTGGAGACCCTCTTCA
AACTATTCAGCCTTCTGGACGAAGACAAGGATGGCATGGTTCAGCTCTCTCTGGCCGAGTGGCTGT
GCTGCGTGTTGGTCTGA

>SGPR_210_SEQ ID NO:10
ATGGCATCCAGCAGTGGGAGGGTCACCATCCAGCTCGTGGATGAGGAGGCTGGGGTCGGAGCCGG
GCGCCTGCAGCTTTTTCGGGGCCAGAGCTATGAGGCAATTCGGGCAGCCTGCCTGGATTCGGGGAT
CCTGTTCCGCGACCCTTACTTCCCTGCTGGCCCTGATGCCCTTGGCTATGACCAGCTGGGGCCGGA
CTCGGAGAAGGCCAAAGGCGTGAAATGGATGAGGCCCCATGAGTTCTGTGCTGAGCCGAAGTTCA
TCTGTGAAGACATGAGCCGCACAGACGTGTGTCAGGGGAGCCTGGGTAACTGCTGGTTCCTTGCA
GCCGCCGCCTCCCTTACTCTGTATCCCCGGCTCCTGCGCCGGGTGGTCCCTCCTGGACAGGATTTCC
AGCATGGCTACGCAGGCGTCTTCCACTTCCAGCTCTGGCAGTTTGGCCGCTGGATGGACGTCGTGG
TGGATGACAGGCTGCCCGTGCGTGAGGGGAAGCTGATGTTCGTGCGCTCGGAACAGCGGAATGAG
TTCTGGGCCCCACTCCTGGAGAAGGCCTACGCCAAGCTCCACGGCTCCTATGAGGTGATGCGGGGC
GGCCACATGAATGAGGCTTTTGTGGATTTCACAGGCGGCGTGGGCGAGGTGCTCTATCTGAGACA
AAACAGCATGGGGCTGTTCTCTGCCCTGCGCCATGCCCTGGCCAAGGAGTCCCTCGTGGGCGCCAC
TGCCCAGAGTGATCGGGGTGAGTACCGCACAGAAGAGGGCCTGGTAAAGGGACACGCGTATTCCA
TCACGGGCACACACAAGGTGTTCCTGGGCTTCACCAAGGTGCGGCTGCTGCGGCTGCGGAACCCA
TGGGGCTGCGTGGAGTGGACGGGGGCCTGGAGCGACAGCTGCCCACGCTGGGACACACTCCCCAC
CGAGTGCCGCGATGCCCTGCTGGTGAAAAAGGAGGATGGCGAGTTCTGGATGGAGCTGCGGGACT
TCCTCCTCCATTTCGACACCGTGCAGATCTGCTCGCTGAGCCCGGAGGTGCTGGGCCCCAGCCCGG
AGGGGGGCGGCTGGCACGTCCACACCTTCCAAGGCCGCTGGGTGCGTGGCTTCAACTCCGGCGGG
AGCCAGCCTAATGCTGAAACCTTCTGGACCAATCCTCAGTTCCGTTTAACGCTGCTGGAGCCTGAT
GAGGAGGATGACGAGGATGAGGAAGGGCCCTGGGGGGGCTGGGGGGCTGCAGGGGCACGGGGCC
CAGCGCGGGGGGGCCGCACGCCCAAGTGCACGGTCCTTCTGTCCCTCATCCAGCGCAACCGGCGG
CGCCTGAGAGCCAAGGGCCTCACTTACCTCACCGTTGGCTTCCACGTGTTCCAGGCAGAGGGCTCC

Figure 1E

ACAGGCACAGACAACGAGCGGACACACGGCTTCACCGGACACAGAGGAGCACAGCTCGCCGGTC
ACACACACGGCCCACAAGAGGCGAGCAAAAGATACACGCAGAACAGCGCTGAGGTAGCCCCAGA
TAGGGAAGCGGACGACGACGGGGGACAGGGGTTCGGCGACGGGCCATGGGAGATCGACGACGTG
ATCAGCGCAGACCTGCAGTCTCTCCAGGGCCCCTACCTGCCCCTGGAGCTGGGGTTGGAGCAGCTG
TTTCAGGAGCTGGCTGGAGAGGAGGAAGAACTCAATGCCTCTCAGCTCCAGGCCTTACTAAGCATT
GCCCTGGAGCCTGCCAGGGCCCATACCTCCACCCCCAGAGAGATCGGGCTCAGGACCTGTGAGCA
GCTGCTGCAGTGTTTCGGGCATGGGCAAAGCCTGGCCTTACACCACTTCCAGCAGCTCTGGGGCTA
CCTCCTGGAGTGGCAGGCCATATTCAACAAGTTCGATGAGGACACCTCTGGAACCATGAACTCCTA
CGAGCTGAGGCTGGCACTGAATGCAGCAGGCTTCCACCTGAACAACCAGCTGACCCAGACCCTCA
CCAGCCGCTACCGGGATAGCCGTCTGCGTGTGGACTTCGAGCGGTTCGTGTCCTGTGTGGCCCACC
TCACCTGCATCTTCTGCCACTGCAGCCAGCACCTGGATGGGGGTGAGGGGGTCATCTGCCTGACCC
ACAGACAGTGGATGGAGGTGGCCACCTTCTCCTAG

>SGPR_290_SEQ ID NO:11
ATGTCTCTGTGGCCACCTTTCCGATGCAGATGGAAGCTGGCGCCAAGGTACTCTAGGAGGGCGTCT
CCACAGCAACCCCAACAGGACTTTGAGGCCCTGCTGGCAGAGTGCCTGAGGAATGGCTGCCTCTTT
GAAGACACCAGCTTCCCGGCCACCCTGAGCTCCATCGGCAGTGGCTCCCTGCTGCAGAAGCTGCCA
CCCCGCCTGCAGTGGAAGAGGCCCCCGGAGCTGCACAGCAATCCCCAGTTTTATTTTGCCAAGGCC
AAAAGGCTGGATCTGTGCCAGGGGATAGTAGGAGACTGCTGGTTCTTGGCTGCTTTGCAAGCTCTG
GCCTTGCACCAGGACATCCTGAGCCGGGTTGTTCCCCTGAATCAGAGTTTCACTGAGAAGTATGCT
GGCATCTTCCGGTTCTGGTTCTGGCACTATGGGAACTGGGTTCCTGTGGTGATCGATGACCGTCTG
CCTGTGAATGAGGCTGGCCAGCTGGTCTTTGTCTCCTCCACCTATAAGAACTTGTTCTGGGGAGCA
CTTCTGGAAAAGGCCTATGCCAAGCTCTCTGGTTCCTATGAAGACTTGCAGTCAGGACAGGTGTCT
GAAGCCCTTGTAGACTTCACTGGAGGGGTGACAATGACCATCAACCTGGCAGAAGCCCATGGCAA
CCTCTGGGACATCCTCATCGAAGCCACCTACAACAGAACCCTCATTGGCTGCCAGACCCACTCAGG
GGAGAAGATTCTGGAGAATGGGCTGGTGGAAGGCCATGCCTATACTCTCACAGGAATCAGGAAGG
TGACCTGCAAACATAGACCTGAATATCTCGTCAAGCTACGGAACCCCTGGGGAAAGGTGGAATGG
AAAGGAGACTGGAGTGACAGTTCAAGTAAATGGGAGCTGCTGAGCCCCAAGGAGAAGATTCTGCT
TCTGAGGAAAGACAATGACGGAGAATTCTGGATGACGCTGCAGGACTTTAAAACACATTTCGTGC
TCCTGGTTATCTGTAAACTGACCCCAGGCCTGTTGAGCCAGGAGGCGGCCCAGAAGTGGACGTAC
ACCATGCGGGAGGGGAGATGGGAGAAGCGGAGCACAGCTGGTGGCCAGAGGCAGTTGCTGCAGG
ACACATTTTGGAAGAACCCGCAGTTCCTGCTGTCTGTCTGGAGGCCCGAGGAGGGCAGGAGATCC
CTGAGGCCCTGCAGCGTGCTGGTGTCCCTGCTCCAGAAGCCCAGGCACAGGTGCCGCAAGCGGAA
GCCTCTCCTCGCCATTGGCTTCTACCTCTATAGGATGAACAAGTACCATGATGACCAGAGGAGACT
GCCCCCTGAGTTCTTCCAGAGAAACACTCCTCTGAGCCAGCCTGATAGGTTTCTCAAGGAGAAAGA
AGTGAGTCAGGAGCTGTGTCTGGAACCAGGGACGTACCTCATCGTGCCTGCATATTGGAGGCCCA
CCAGAAGTCAGAGTTCGTCCTCAGGGTCTTCTCCAGGAAGCACATCTTTTATGAAATTGGCAGCAA
TTCTGGTGTCGTCTTCTCAAAGGAGATANAAGACCAAAATGAAAGGCAGGATGAATTCTTCACCA
AATTCTTTTGNAAAGCATCCAGAGATTAATGCAGTTCAACTTCAGAACCTCCTGNACCAGATGACC
TGGTCAAGTCTGGGGAGCAGACAGCCCTTCTTTAGCCTGGAAGCCTGCCAGGGGATCCTGGCCTTA
CTGGACGTATCCTTTCAGCTTAATGCATCAGGTACTATGAGCATCCAGGAATTCAGGGACCTGTGG
AAGCAGCTGAAGCTCTCTCAGAAGGTTTTCCACAAGCAAGACCGTGGGTCAGGATACCTGAACTG
GGAGCAGCTGCACGCTGCCATGAGGGAGGCAGGAATCATGCTCAGTGATGACGTCTGTCAGCTGA
TGCTCATCCGCTACGGCGGCCCCCGCCTCCAGATGGACTTTGTCAGTTTCATCCACTTGATGCTGCG
TGTAGAGAACATGGAGGGTAAGCTGGCGGGAAGCTGGGGAGGGCCAGGTCTTCCTCTGCTGCCCC
ATGACTTCCCACCTGTCCCTAGTTTAAGCACAAGGGAGGACAGCCGCCATCCCAGAAACAGCAGA
CCAGGGAAGCTGTGGGGACCTCCAGCCAAGTGCCTGTGA

>SGPR_116_SEQ ID NO:12
ATGGTGGCTCACATAAACAACAGCCGGCTCAAGGCCAAGGGCGTGGGCCAGCACGACAACGCCCA
GAACTTTGGTAACCAGAGCTTTGAGGAGCTGCGAGCAGCCTGTCTAAGAAAGGGGGAGCTCTTCG
AGGACCCCTTATTCCCTGCTGAACCCAGCTCACTGGGCTTCAAGGACCTGGGCCCCAACTCCAAAA
ATGTGCAGAACATCTCCTGGCAGCGGCCCAAGGATATCATAAACAACCCTCTATTCATCATGGATG
GGATTTCTCCAACAGACATCTGCCAGGGGATCCTCGGGGACTGCTGGCTGCTGGCTGCCATCGGCT
CCCTTACCACCTGCCCCAAACTGCTATACCGCGTGGTGCCCAGAGGACAGAGCTTCAAGAAAAAC
TATGCTGGCATCTTCCATTTTCAGATTTGGCAGTTTGGACAGTGGGTGAACGTGGTGGTAGATGAC
CGGCTGCCCACAAAGAATGACAAGCTGGTGTTTGTGCACTCAACCGAACGCAGTGAGTTCTGGAG
TGCCCTGCTGGAGAAGGCGTATGCCAAGCTGAGTGGGTCCTATGAAGCATTGTCAGGGGGCAGTA

Figure 1F

CCATGGAGGGCCTTGAGGACTTCACAGGAGGCGTGGCCCAGAGCTTCCAACTCCAGAGGCCCCCT
CAGAACCTGCTCAGGCTCCTTAGGAAGGCCGTGGAGCGATCCTCCCTCATGGGTTGCTCCATTGAA
GTCACCAGTGATAGTGAACTGGAATCCATGACTGACAAGATGCTGGTGAGAGGGCACGCTTACTC
TGTGACTGGCCTTCAGGATGTCCACTACAGAGGCAAAATGGAAACACTGATTCGGGTCCGGAATC
CCTGGGGCCGGATTGAGTGGAATGGAGCTTGGAGTGACAGTGCCAGGGAGTGGGAAGAGGTGGC
CTCAGACATCCAGATGCAGCTGCTGCACAAGACGGAGGACGGGGAGTTCTGGATGTCCTACCAAG
ATTTCCTGAACAACTTCACGCTCCTGGAGATCTGCAACCTCACGCCTGATACACTCTCTGGGGACT
ACAAGAGCTACTGGCACACCACCTTCTACGAGGGCAGCTGGCGCAGAGGCAGCTCCGCAGGGGGC
TGCAGGAACCACCCTGGCACGTTCTGGACCAACCCCCAGTTTAAGATCTCTCTTCCTGAGGGGGAT
GACCCAGAGGATGACGCAGAGGGCAATGTTGTGGTCTGCACCTGCCTGGTGGCCCTAATGCAGAA
GAACTGGCGGCATGCACGGCAGCAGGGAGCCCAGCTGCAGACCATTGGCTTTGTCCTCTACGCGG
TCCCAAAAGAGTTTCAGAACATTCAGGATGTCCACTTGAAGAAGGAATTCTTCACGAAGTATCAG
GACCACGGCTTCTCAGAGATCTTCACCAACTCACGGGAGGTGAGCAGCCAACTCCGGCTGCCTCCG
GGGGAATATATCATTATTCCCTCCACCTTTGAGCCACACAGAGATGCTGACTTCCTGCTTCGGGTCT
TCACCGAGAAGCACAGCGAGTCATGGGAATTGGATGAAGTCAACTATGCTGAGCAACTCCAAGAG
GAAAAGGTCTCTGAGGATGACATGGACCAGGACTTCCTACATTTGTTTAAGATAGTGGCAGGAGA
GGGCAAGGAGATAGGGGTGTATGAGCTCCAGAGGCTGCTCAACAGGATGGCCATCAAATTCAAAA
GCTTCAAGACCAAGGGCTTTGGCCTGGATGCTTGCCGCTGCATGATCAACCTCATGGATAAAGATG
GCTCTGGCAAGCTGGGGCTTCTAGAGTTCAAGATCCTGTGGAAAAAACTCAAGAAATGGATGGAC
ATCTTCAGAGAGTGTGACCAGGACCATTCAGGCACCTTGAACTCCTATGAGATGCGCCTGGTTATT
GAGAAAGCAGGCATCAAGCTGAACAACAAGGTAATGCAGGTCCTGGTGGCCAGGTATGCAGATG
ATGACCTGATCATAGACTTTGACAGCTTCATCAGCTGTTTCCTGAGGCTAAAGACCATGTTCACAT
TCTTTCTAACCATGGACCCCAAGAATACTGGCCATATTTGCTTGAGCCTGGAACAGTGGCTGCAGA
TGACCATGTGGGGATAG

>SGPR_003_SEQ ID NO:13
ATGCGGGCGGGCCGGGGCGCGACGCCGGCGAGGGAGCTGTTCCGGGACGCCGCCTTCCCCGCCGC
GGACTCCTCGCTCTTCTGCGACTTGTCTACGCCGCTGGCCCAGTTCCGCGAGGACATCACGTGGAG
GCGGCCCCAGGAGATTTGTGCCACACCCCGGCTGTTTCCAGATGACCCACGGGAAGGGCAGGTGA
AGCAGGGGCTGCTGGGGGATTGCTGGTTCCTGTGTGCCTGCGCCGCGCTGCAGAAGAGCAGGCAC
CTCCTGGACCAGGTCATTCCTCCGGGACAGCCGAGCTGGGCCGACCAGGAGTACCGGGGCTCCTTC
ACCTGTCGCATTTGGCAGTTTGGACGCTGGGTGGAGGTGACCACAGATGACCGCCTGCCGTGCCTT
GCAGGGAGACTCTGTTTCTCCCGCTGCCAGAGGGAGGATGTGTTCTGGCTCCCCTTACTGGAAAAG
GTCTACGCCAAGGTCCATGGGTCCTACGAGCACCTGTGGGCCGGGCAGGTGGCGGATGCCCTGGT
GGACCTGACCGGCGGCCTGGCAGAAAGATGGAACCTGAAGGGCGTAGCAGGAAGCGGAGGCCAG
CAGGACAGGCCAGGCCGCTGGGAGCACAGGACTTGTCGGCAGCTGCTCCACCTGAAGGACCAGTG
TCTGATCAGCTGCTGCGTGCTCAGCCCCAGAGCAGGTGCCCGGGAGCTGGGGGAGTTCCATGCCTT
CATTGTCTCGGACCTGCGGGAGCTCCAGGGTCAGGCGGGCCAGTGCATCCTGCTGCTGCGGATCCA
GAACCCCTGGGGCCGGCGGTGCTGGCAGGGGCTCTGGAGAGAGGGGGGTGAAGGGTGGAGCCAG
GTAGATGCAGCGGTAGCATCTGAGCTCCTGTCCCAGCTCCAGGAAGGGGAGTTCTGGGTGGAGGA
GGAGGAGTTCCTCAGGGAGTTTGACGAGCTCACCGTTGGCTACCCGGTCACGGAGGCCGGCCACC
TGCAGAGCCTCTACACAGAGAGGCTGCTCTGCCATACGCGGGCGCTGCCTGGGGCCTGGGTCAAG
GGCCAGTCAGCAGGAGGCTGCCGGAACAACAGCGGCTTTCCCAGCAACCCCAAATTCTGGCTGCG
GGTCTCAGAACCGAGTGAGGTGTACATTGCCGTCCTGCAGAGATCCAGGCTGCACGCGGCGGACT
GGGCAGGCCGGGCCCGGGCACTGGTGGGTGACAGTCATACTTCGTGGAGCCCAGCGAGCATCCCG
GGCAAGCACTACCAGGCTGTGGGTCTGCACCTCTGGAAGGTAGAGAAGCGGCGGGTCAATCTGCC
TAGGGTCCTGTCCATGCCCCCCGTGGCTGGCACCGCGTGCCATGCATACGACCGGGAGGTCCACCT
GCGTTGTGAGCTCTCACCGGGCTACTACCTGGCTGTCCCCAGCACCTTCCTGAAGGACGCGCCAGG
GGAGTTCCTGCTCCGAGTCTTCTCTACCGGGCGAGTCTCCCTTAGGTCCCAGAGGGTGGAAGGAGC
CAGGACGCACCCCCACTGCTGCTGCAGGAGCCGCTGCTGA

>SGPR_016_SEQ ID NO:14
ATGTTCCTTCTTCTGGTGCTTCTCACTGGACTTGGTGGGATGCATGCAGACCTCAATCCTCATAAAA
TCTTCCTACAGACCACAATTCCAGAGAAGATTTCATCATCGGATGCAAAAACAGATCCAGAACAT
AATGTAATTTTAATAATATTTTTACTAGAAATCATGTTTTTATTATTTTTGCCTAGATCAATTTTATC
TTCAGCTTCTGTTATTAATTCTTATGACGAAAATGACATCCGTCATTCCAAACCTCTGCTAGTTCAG
ATGGATTGCATTTATAATGGATATGTTGCGGGTATTCCAAATTCTCTTGTGACTCTCAGCGTATGTT
CAGGACTCAGGTTGGGAACAATGCAGCTGAAAAACATCTCATATGGAATTGAACCGATGGAGGCT

Figure 1G

AAAACTGACTTTATTAAGTTATTCCCTCGATATATTGAAATGCATATTGTTGTGGACAAAAATTTG
GTAAAAACAATAAAAAGTATCTGGNNNATGTTTTCTCAGCTTAAAACAAGTATTACGCTATCTTCT
TTGGAGCTCTGGTCAGATGAAAATAAGATTTCAACTAATGGGGTTGCTGATGATGTACTACAAAGG
TTTTTATCATGGAAACAAAAATTTATGTCTCAAAAGTCCAATATCGTGGCATATTTATTAATGNNN
TACTCTGGTGGTGTAAAGGATTTTAACATCTGTAGCTTGGATGACTTTAAATATATTTCTTCTCATA
ATGGCCTTACATGTCTTCAGACAAACCCTCTTGAAATGCCAACCTACACACACAGGAGAATATGTG
GCAATGGGTTGTTGGAAGGAAGTGAAGAATGTGACTGTGGCACTAAAGAC

>SGPR_352_SEQ ID NO:15
ATGGCTCCCGCCTGCCAGATCCTCCGCTGGGCCCTCGCCCTGGGGCTGGGCCTCATGTTCGAGGTC
ACGCACGCCTTCCGGTCTCAAGATGAGTTCCTGTCCAGTCTGGAGAGCTATGAGATCGCCTTCCCC
ACCCGCGTGGACCACAACGGGGCACTGCTGGCCTTCTCGCCACCTCCTCCCCGGAGGCAGCGCCGC
GGCACGGGGGCCACAGCCGAGTCCCGCCTCTTCTACAAAGTGGCCTCGCCCAGCACCCACTTCCTG
CTGAACCTGACCCGCAGCTCCCGTCTACTGGCAGGGCACGTCTCCGTGGAGTACTGGACACGGGA
GGGCCTGGCCTGGCAGAGGGCGGCCCGGCCCCACTGCCTCTACGCTGGTCACCTGCAGGGCCAGG
CCAGCAGCTCCCATGTGGCCATCAGCACCTGTGGAGGCCTGCACGGCCTGATCGTGGCAGACGAG
GAAGAGTACCTGATTGAGCCCCTGCACGGTGGGCCCAAGGGTTCTCGGAGCCCGGAGGAAAGTGG
ACCACATGTGGTGTACAAGCGTTCCTCTCTGCGTCACCCCCACCTGGACACAGCCTGTGGAGTGAG
AGATGAGAAACCGTGGAAAGGGCGGCCATGGTGGCTGCGGACCTTGAAGCCACCGCCTGCCAGGC
CCCTGGGGAATGAAACAGAGCGTGGCCAGCCAGGCCTGAAGCGATCGGTCAGCCGAGAGCGCTAC
GTGGAGACCCTGGTGGTGGCTGACAAGATGATGGTGGCCTATCACGGGCGCCGGGATGTGGAGCA
GTATGTCCTGGCCGTCATGAACATTGTTGCCAAACTTTTCCAGGACTCGAGTCTGGGAAGCACCGT
TAACATCCTCGTAACTCGCCTCATCCTGCTCACGGAGGACCAGCCCACTCTGGAGATCACCCACCA
TGCCGGGAAGTCCCTGGACAGCTTCTGTAAGTGGCAGAAATCCATCGTGAACCACAGCGGCCATG
GCAATGCCATTCCAGAGAACGGTGTGGCTAACCATGACACAGCAGTGCTCATCACACGCTATGAC
ATCTGCATCTACAAGAACAAACCCTGCGGCACACTAGGCCTGGCCCCGGTGGGCGGAATGTGTGA
GCGCGAGAGAAGCTGCAGCGTCAATGAGGACATTGGCCTGGCCACAGCGTTCACCATTGCCCACG
AGATCGGGCACACATTCGGCATGAACCATGACGGCGTGGGAAACAGCTGTGGGGCCCGTGGTCAG
GACCCAGCCAAGCTCATGGCTGCCCACATTACCATGAAGACCAACCCATTCGTGTGGTCATCCTGC
AGCCGTGACTACATCACCAGCTTTCTAGACTCGGGCCTGGGGCTCTGCCTGAACAACCGGCCCCCC
AGACAGGACTTTGTGTACCCGACAGTGGCACCGGGCCAAGCCTACGATGCAGATGAGCAATGCCG
CTTTCAGCATGGAGTCAAATCGCGTCAGTGTAAATACGGGGAGGTCTGCAGCGAGCTGTGGTGTCT
GAGCAAGAGCAACCGGTGCATCACCAACAGCATCCCGGCCGCCGAGGGCACGCTGTGCCAGACGC
ACACCATCGACAAGGGGTGGTGCTACAAACGGGTCTGTGTCCCCTTTGGGTCGCGCCCAGAGGGT
GTGGACGGAGCCTGGGGGCCGTGGACTCCATGGGGCGACTGCAGCCGGACCTGTGGCGGCGGCGT
GTCCTCTTCTAGCCGTCACTGCGACAGCCCCAGGCCAACCATCGGGGGCAAGTACTGTCTGGGTGA
GAGAAGGCGGCACCGCTCCTGCAACACGGATGACTGTCCCCCTGGCTCCCAGGACTTCAGAGAAG
TGCAGTGTTCTGAATTTGACAGCATCCCTTTCCGTGGGAAATTCTACAAGTGGAAAACGTACCGGG
GAGGGGGCGTGAAGGCCTGCTCGCTCACGTGCCTAGCGGAAGGCTTCAACTTCTACACGGAGAGG
GCGGCAGCCGTGGTGGACGGGACACCCTGCCGTCCAGACACGGTGGACATTTGCGTCAGTGGCGA
ATGCAAGCACGTGGGCTGCGACCGAGTCCTGGGCTCCGACCTGCGGGAGGACAAGTGCCGAGTGT
GTGGCGGTGACGGCAGTGCCTGCGAGACCATCGAGGGCGTCTTCAGCCCAGCCTCACCTGGGGCC
GGGTACGAGGATGTCGTCTGGATTCCCAAAGGCTCCGTCCACATCTTCATCCAGGATCTGAACCTC
TCTCTCAGTCACTTGGCCCTGAAGGGAGACCAGGAGTCCCTGCTGCTGGAGGGGCTGCCCGGGAC
CCCCCAGCCCCACCGTCTGCCTCTAGCTGGGACCACCTTTCAACTGCGACAGGGGCCAGACCAGGT
CCAGAGCCTCGAAGCCCTGGGACCGATTAATGCATCTCTCATCGTCATGGTGCTGGCCCGGACCGA
GCTGCCTGCCCTCCGCTACCGCTTCAATGCCCCCATCGCCCGTGACTCGCTGCCCCCCTACTCCTGG
CACTATGCGCCCTGGACCAAGTGCTCGGCCCAGTGTGCAGGCGGTAGCCAGGTGCAGGCGGTGGA
GTGCCGCAACCAGCTGGACAGCTCCGCGGTCGCCCCCCACTACTGCAGTGCCCACAGCAAGCTGC
CCAAAAGGCAGCGCGCCTGCAACACGGAGCCTTGCCCTCCAGACTGGGTTGTAGGGAACTGGTCG
CTCTGCAGCCGCAGCTGCGATGCAGGCGTGCGCAGCCGCTCGGTCGTGTGCCAGCGCCGCGTCTCT
GCCGCGGAGGAGAAGGCGCTGGACGACAGCGCATGCCCGCAGCCGCGCCCACCTGTACTGGAGGC
CTGCCACGGCCCCACTTGCCCTCCGGAGTGGGCGGCCCTCGACTGGTCTGAGTGCACCCCCAGCTG
CGGGCCGGGCCTCCGCCACCGCGTGGTCCTTTGCAAGAGCGCAGACCACCGCGCCACGCTGCCCC
CGGCGCACTGCTCACCCGCCGCCAAGCCACCGGCCACCATGCGCTGCAACTTGCGCCGCTGCCCCC
CGGCCCGCTGGGTGGCTGGCGAGTGGGGTGAGTGCTCTGCACAGTGCGGCGTCGGGCAGCGGCAG
CGCTCGGTGCGCTGCACCAGCCACACGGGCCAGGCGTCGCACGAGTGCACGGAGGCCCTGCGGCC
GCCCACCACGCAGCAGTGTGAGGCCAAGTGCGACAGCCCAACCCCCGGGGACGGCCCTGAAGAGT

Figure 1H

GCAAGGATGTGAACAAGGTCGCCTACTGCCCCCTGGTGCTCAAATTTCAGTTCTGCAGCCGAGCCT
ACTTCCGCCAGATGTGCTGCAAAACCTGCCAGGGCCACTAG

>SGPR_050_SEQ ID NO:16
ATGAAGCCCCGCGCGCGCGGATGGCGGGGCTTGGCGGCGCTGTGGATGCTGCTGGCGCAGGTGGC
CGAGCAGGCACCTGCGTGCGCCATGGGACCCGCAGCGGCAGCGCCTGGGAGCCCGAGCGTCCCGC
GTCCTCCTCCACCCGCGGAGCGGCCGGGCTGGATGGAAAAGGGCGAATATGACCTGGTCTCTGCC
TACGAGGTTGACCACAGGGGCGATTACGTGTCCCATGAAATCATGCACCATCAGCGGCGGAGAAG
AGCAGTGGCCGTGTCCGAGGTTGAGTCTCTTCACCTTCGGCTGAAAGGCCCCAGGCACGACTTCCA
CATGGATCTGAGGACTTCCAGCAGCCTAGTGGCTCCTGGCTTTATTGTGCAGACGTTGGGAAAGAC
AGGCACTAAGTCTGTGCAGACTTTACCGCCAGAGGACTTCTGTTTCTATCAAGGCTCTTTGCGATC
ACACAGAAACTCCTCAGTGGCCCTTTCAACCTGCCAAGGCTTGTCAGGCATGATACGAACAGAAG
AGGCAGATTACTTCCTAAGGCCACTTCCTTCACACCTCTCATGGAAACTCGGCAGAGCTGCCCAAG
GCAGCTCGCCATCCCACGTACTGTACAAGAGATCCACAGAGCCCCATGCTCCTGGGGCCAGTGAG
GTCCTGGTGACCTCAAGGACATGGGAGCTGGCACATCAACCCCTGCACAGCAGCGACCTTCGCCT
GGGACTGCCACAAAAGCAGCATTTCTGTGGAAGACGCAAGAAATACATGCCCCAGCCTCCCAAGG
AAGACCTCTTCATCTTGCCAGATGAGTATAAGTCTTGCTTACGGCATAAGCGCTCTCTTCTGAGGTC
CCATAGAAATGAAGAACTGAACGTGGAGACCTTGGTGGTGGTCGACAAAAAGATGATGCAAAACC
ATGGCCATGAAAATATCACCACCTACGTGCTCACGATACTCAACATGGTATCTGCTTTATTCAAAG
ATGGAACAATAGGAGGAAACATCAACATTGCAATTGTAGGTCTGATTCTTCTAGAAGATGAACAG
CCAGGACTGGTGATAAGTCACCACGCAGACCACACCTTAAGTAGCTTCTGCCAGTGGCAGTCTGG
ATTGATGGGGAAAGATGGGACTCGTCATGACCACGCCATCTTACTGACTGGTCTGGATATATGTTC
CTGGAAGAATGAGCCCTGTGACACTTTGGGATTTGCACCCATAAGTGGAATGTGTAGTAAATATCG
CAGCTGCACGATTAATGAAGATACAGGTCTTGGACTGGCCTTCACCATTGCCCATGAGTCTGGACA
CAACTTTGGCATGATTCATGATGGAGAAGGGAACATGTGTAAAAAGTCCGAGGGCAACATCATGT
CCCCTACATTGGCAGGACGCAATGGAGTCTTCTCCTGGTCACCCTGCAGCCGCCAGTATCTACACA
AATTTCTAAGCACCGCTCAAGCTATCTGCCTTGCTGATCAGCCAAAGCCTGTGAAGGAATACAAGT
ATCCTGAGAAATTGCCAGGAGAATTATATGATGCAAACACACAGTGCAAGTGGCAGTTCGGAGAG
AAAGCCAAGCTCTGCATGCTGGACTTTAAAAAGGACATCTGTAAAGCCCTGTGGTGCCATCGTATT
GGAAGGAAATGTGAGACTAAATTTATGCCAGCAGCAGAAGGCACAATTTGTGGGCATGACATGTG
GTGCCGGGGAGGACAGTGTGTGAAATATGGTGATGAAGGCCCCAAGCCCACCCATGGCCACTGGT
CGGACTGGTCTTCTTGGTCCCCATGCTCCAGGACCTGCGGAGGGGGAGTATCTCATAGGAGTCGCC
TCTGCACCAACCCCAAGCCATCGCATGGAGGGAAGTTCTGTGAGGGCTCCACTCGCACTCTGAAGC
TCTGCAACAGTCAGAAATGTCCCCGGGACAGTGTTGACTTCCGTGCTGCTCAGTGTGCCGAGCACA
ACAGCAGACGATTCAGAGGGCGGCACTACAAGTGGAAGCCTTACACTCAAGTAGAAGATCAGGAC
TTATGCAAACTCTACTGTATCGCAGAAGGATTTGATTTCTTCTTTTCTTTGTCAAATAAAGTCAAAG
ATGGGACTCCATGCTCGGAGGATAGCCGTAATGTTTGTATAGATGGGATATGTGAGAGAGTTGGA
TGTGACAATGTCCTTGGATCTGATGCTGTTGAAGACGTCTGTGGGGTGTGTAACGGGAATAACTCA
GCCTGCACGATTCACAGGGGTCTCTACACCAAGCACCACCACACCAACCAGTATTATCACATGGTC
ACCATTCCTTCTGGAGCCCGGAGTATCCGCATCTATGAAATGAACGTCTCTACCTCCTACATTTCTG
TGCGCAATGCCCTCAGAAGGTACTACCTGAATGGGCACTGGACCGTGGACTGGCCCGGCCGGTAC
AAATTTTCGGGCACTACTTTCGACTACAGACGGTCCTATAATGAGCCCGAGAACTTAATCGCTACT
GGACCAACCAACGAGACACTGATTGTGGAGCTGCTGTTTCAGGGAAGGAACCCGGGTGTTGCCTG
GGAATACTCCATGCCTCGCTTGGGGACCGAGAAGCAGCCCCCTGCCCAGCCCAGCTACACTTGGG
CCATCGTGCGCTCTGAGTGCTCCGTGTCCTGCGGAGGGGGACAGATGACCGTGAGAGAGGGCTGC
TACAGAGACCTGAAGTTTCAAGTAAATATGTCCTTCTGCAATCCCAAGACACGACCTGTCACGGGG
CTGGTGCCTTGCAAAGTATCTGCCTGTCCTCCCAGCTGGTCCGTGGGGAACTGGAGTGCCTGCAGT
CGGACGTGTGGCGGGGGTGCCCAGAGCCGCCCCGTGCAGTGCACACGGCGGGTGCACTATGACTC
GGAGCCAGTCCCGGCCAGCCTGTGCCCTCAGCCTGCTCCCTCCAGCAGGCAGGCCTGCAACTCTCA
GAGCTGCCCACCTGCATGGAGCGCCGGGCCCTGGGCAGAGTGCTCACACACCTGTGGGAAGGGGT
GGAGGAAGCGGGCAGTGGCCTGTAAGAGCACCAACCCCTCGGCCAGAGCGCAGCTGCTGCCCGAC
GCTGTCTGCACCTCCGAGCCCAAGCCCAGGATGCATGAAGCCTGTCTGCTTCAGCGCTGCCACAAG
CCCAAGAAGCTGCAGTGGCTGGTGTCCGCCTGGTCCCAGTGCTCTGTGACATGTGAAAGAGGAAC
ACAGAAAAGATTCTTAAAATGTGCTGAAAAGTATGTTTCTGGAAAGTATCGAGAGCTGGCCTCAA
AGAAGTGCTCACATTTGCCGAAGCCCAGCCTGGAGCTGGAACGTGCCTGCGCCCCGCTTCCATGCC
CCAGGCACCCCCCATTTGCTGCTGCGGGACCCTCGAGGGGCAGCTGGTTTGCCTCACCCTGGTCTC
AGTGCACGGCCAGCTGTGGGGGAGGCGTTCAGACGAGGTCCGTGCAGTGCCTGGCTGGGGGCCGG
CCGGCCTCAGGCTGCCTCCTGCACCAGAAGCCTTCGGCCTCCCTGGCCTGCAACACTCACTTCTGC

Figure 1I

CCCATTGCAGAGAAGAAAGATGCCTTCTGCAAAGACTACTTCCACTGGTGCTACCTGGTACCCCAG
CACGGGATGTGCAGCCACAAGTTCTACGGCAAGCAGTGCTGCAAGACTTGCTCTAAGTCCAACTTG
TGA

>SGPR_282_SEQ ID NO:17
ATGAGGCAGGCAGAGGCGCGGGTCACCCTTAGGGCCCCCCTCTTGCTGCTGGGGCTCTGGGTGCTC
CTGACTCCAGTCCGGTGTTCTCAAGGCCATCCCTCGTGGCACTACGCATCCTCCAAGGTGGTGATT
CCCAGGAAGGAGACGCACCACGGCAAAGACCTTCAGTTTCTGGGCTGGCTGTCCTACAGCCTGCA
TTTTGGGGGTCAAAGACACATCATTCACATGCGGAGGAAACACCTTCTTTGGCCCAGACATCTGCT
GGTGACAACTCAGGATGACCAAGGAGCCTTGCAGATGGATGACCCCTACATCCCTCCAGACTGCT
ACTATCTCAGCTACCTGGAGGAGGTTCCTCTGTCCATGGTCACCGTGGACATGTGCTGTGGGGGCC
TCAGAGGCATCATGAAGCTGGACGACCTTGCCTATGAAATCAAACCCCTCCAGGATTCCCGCAGG
CTTGAACATGTTTCTCAGATAGTGGCCGAGCCCAACGCAACGGGGCCCACATTTAGAGATGGTGA
CAATGAGGAGACAAACCCCCTGTTCTCTGAAGCAAATGACAGCATGAATCCCAGGATATCTAATT
GGCTGTATAGTTCTCATAGAGGCAATATAAAAGGCCACGTTCAATGTTCCAATTCATATTGTCGTG
TAGATGACAATATTACAACTTGTTCCAAGGAGGTGGTCCAGATGTTCAGTCTCAGTGACAGCATTG
TTCAAAATATTGATCTGCGGTACTATATTTATCTTTTGACCATATATAATAATTGTGACCCAGCCCC
TGTGAATGACTATCGAGTTCAGAGTGCAATGTTTACCTATTTTAGAACAACCTTTTTTGATACTTTT
CGTGTTCATTCACCCACACTACTTATTAAAGAGGCACCACATGAATGTAACTATGAACCACAAAGG
TATAGCTTCTGTACACATTTAGGCCTATTACACATTGGTACTCTAGGCAGACATTATTTATTAGTAG
CCGTCATAACAACCCAGACACTGATGAGAAGTACTGGTGAGAAGTACGATGATAACTACTGCACA
TGTCAGAAAAGGGCCTTCTGCATTATGCAGCAATATCCTGGGATGACAGATGCGTTCAGTAACTGT
TCTTATGGACATGCACAAAATTGTTTTGTACATTCAGCCCGGTGTGTTTTCGAAACACTTGCTCCTG
TGTATAATGAAACCATGACAATGGTTCGCTGTGGAAACCTCATAGCGGATGGGAGGGAGGAATGT
GACTGTGGCTCCTTCAAGCAGTGTTATGCCAGTTATTGCTGCCGAAGTGACTGTCGCTTAACACCG
GGGAGCATCTGTCATATAGGAGAGTGCTGTACAAACTGCAGCTACTCCCCACCAGGGACTCTCTGC
AGACCTATCCAAAATATATGTGACCTTCCAGAGTACTGTCACGGGACCACCGTGACATGCCCCGCA
AACTTTTATATGCAAGATGGAACCCCGTGCACTGAAGAAGGCTACTGCTATCATGGGAACTGCACT
GACCGCAATGTGCTCTGCAAGGTAATCTTTGGTGTCAGTGCTGAGGAGGCTCCTGAGGTCTGCTAT
GACATAAATCTTGAAAGTTACCGATTTGGACATTGTACTCGACGACAAACAGCTCTCAACAACCAG
GCTTGTGCAGGAATAGATAAGTTTTGTGGAAGACTGCAGTGTACCAGTGTGACCCATCTTCCCCGG
CTGCAGGAACATGTTTCATTCCATCACTCAGTGACAGGAGGATTTCAGTGTTTTGGACTGGATGAC
CACCGTGCAACAGACACAACTGATGTTGGGTGTGTGATAGATGGCACTCCTTGTGTTCATGGAAAC
TTCTGTAATAACACCAGGTGCAATGCGACTATCACTTCACTGGGCTACGACTGTCGCCCTGAGAAG
TGCAGTCATAGAGGGGTGTGCAACAACAGAAGGAACTGCCATTGCCATATAGGCTGGGATCCTCC
ACTGTGCCTAAGAAGAGGTGCTGGTGGGAGTGTCGACAGCGGGCCACCTCCAAAAATAACACGTT
CGGTCAAACAAAGCCAACAATCAGTGATGTATCTGAGAGTGGTCTTTGGTCGTATTTACACCTTCA
TAATTGCACTGCTCTTTGGGATGGCCACAAATGTGCGAACTATCAGGACCACCACTGTTAAGGGAT
GGACAGTTACTAACCCTGAATAA

>SGPR_046_SEQ ID NO:18
ATGGTGGAAAAGCATGGCAAGGGAAATGTCACCACATACATTCTCACAGTAATGAACATGGTTTC
TGGCCTATTTAAAGATGGGACTATTGGAAGTGACATAAACGTGGTTGTGGTGAGCCTAATTCTTCT
GGAACAAGAACCTGGAGGATTATTGATCAACCATCATGCAGACCAGTCTCTGAATAGTTTTTGTCA
ATGGCAGTCTGCCCTCATTGGAAAGAATGGCAAGAGACATGATCATGCCATCTTACTAACAGGATT
TGATATTTGTTCTTGGAAGAATGAACCATGTGACACTCTAGGGTTTGCCCCCATCAGTGGAATGTG
CTCTAAGTACCGAAGTTGTACCATCAATGAGGACACAGGACTTGGCCTTGCCTTCACCATCGCTCA
TGAGTCAGGGCACAACTTTGGTATGATTCACGACGGAGAAGGGAATCCCTGCAGAAAGGCTGAAG
GCAATATCATGTCTCCCACACTGACCGGAAACAATGGAGTGTTTTCATGGTCTTCCTGCAGCCGCC
AGTATCTCAAGAAATTCCTCAGCACACCTCAGGCGGGGTGTCTAGTGGATGAGCCCAAGCAAGCA
GGACAGTATAAATATCCGGACAAACTACCAGGACAGATTTATGATGCTGACACACAGTGTAAATG
GCAATTTGGAGCAAAAGCCAAGTTATGCAGCCTTGGTTTTGTGAAGGATATTTGCAAATCACTTTG
GTGCCACCGAGTAGGCCACAGGTGTGAGACCAAGTTTATGCCCGCAGCAGAAGGGACCGTTTGTG
GCTTGAGTATGTGGTGTCGGCAAGGCCAGTGCGTAAAGTTTGGGGAGCTCGGGCCCCGGCCCATC
CACGGCCAGTGGTCCGCCTGGTCGAAGTGGTCAGAATGTTCCCGGACATGTGGTGGAGGAGTCAA
GTTCCAGGAGAGACACTGCAATAACCCCAAGCCTCAGTATGGTGGCTTATTCTGTCCAGGTTCTAG
CCGTATTTATCAGCTGTGCAATATTAACCCTTGCAATGAAAATAGCTTGGATTTTCGGGCTCAACA
GTGTGCAGAATATAACAGCAAACCTTTCCGTGGATGGTTCTACCAGTGGAAACCCTATACAAAAGT

Figure 1J

GGAAGAGGAAGATCGATGCAAACTGTACTGCAAGGCTGAGAACTTTGAATTTTTTTTTGCAATGTC
CGGCAAAGTGAAAGATGGAACTCCCTGCTCCCCAAACAAAAATGATGTTTGTATTGACGGGGTTT
GTGAACTAGTGGGATGTGATCATGAACTAGGCTCTAAAGCAGTTTCAGATGCTTGTGGCGTTTGCA
AAGGTGATAATTCAACTTGCAAGTTTTATAAAGGCCTGTACCTCAACCAGCATAAAGCAAATGAAT
ATTATCCGGTGGTCCTCATTCCAGCTGGCGCCCGAAGCATCGAAATCCAGGAGCTGCAGGTTTCCT
CCAGTTACCTCGCAGTTCGAAGCCTCAGTCAAAAGTATTACCTCACCGGGGGCTGGAGCATCGACT
GGCCTGGGGAGTTCCCCTTCGCTGGGACCACGTTTGAATACCAGCGCTCTTTCAACCGCCCGGAAC
GTCTGTACGCGCCAGGGCCCACAAATGAGACGCTGGTCTTTGAAATTCTGATGCAAGGCAAAAAT
CCAGGGATAGCTTGGAAGTATGCACTTCCCAAGGTCATGAATGGAACTCCACCAGCCACAAAAAG
ACCTGCCTATACCTGGAGTATCGTGCAGTCAGAGTGCTCCGTCTCCTGTGGTGGAGGTTACATAAA
TGTAAAGGCCATTTGCTTGCGAGATCAAAATACTCAAGTCAATTCCTCATTCTGCAGTGCAAAAAC
CAAGCCAGTAACTGAGCCCAAAATCTGCAACGCTTTCTCCTGCCCGGCTTACTGGATGCCAGGTGA
ATGGAGTACATGCAGCAAGGCCTGTGCTGGAGGCCAGCAGAGCCGAAAGATCCAGTGTGTGCAAA
AGAAGCCCTTCCAAAAGGAGGAAGCAGTGTTGCATTCTCTCTGTCCAGTGAGCACACCCACTCAG
GTCCAAGCCTGCAACAGCCATGCCTGCCCTCCACAATGGAGCCTTGGACCCTGGTCTCAGTGTTCC
AAGACCTGTGGACGAGGGGTGAGGAAGCGTGAACTCCTCTGCAAGGGCTCTGCCGCAGAAACCCT
CCCCGAGAGGAAGCGTGAACTCCTCTGCAAGGGCTCTGCCGCAGAAACCCTCCCCGAGAGCCAGT
GTACCAGTCTCCCCAGACCTGAGCTGCAGGAGGGCTGTGTGCTTGGACGATGCCCCAAGAACAGC
CGGCTACAGTGGGTCGCTTCTTCGTGGAGCGAGTGTTCTGCAACCTGTGGTTTGGGTGTGAGGAAG
AGGGAGATGAAGTGCAGCGAGAAGGGCTTCCAGGGAAAGCTGATAACTTTCCCAGAGCGAAGAT
GCCGTAATATTAAGAAACCAAATCTGGACTTGGAAGAGACCTGCAACCGACGGGCTTGCCCAGCC
CATCCAGTGTACAACATGGTAGCTGGATGGTATTCATTGCCGTGGCAGCAGTGCACAGTCACCTGT
GGGGGAGGGGTCCAGACCCGGTCAGTCCACTGTGTTCAGCAAGGCCGGCCTTCCTCAAGTTGTCTG
CTCCATCAGAAACCTCCGGTGCTACGAGCCTGTAATACAAACTTCTGTCCAGCTCCTGAAAAGAGA
GAGGATCCATCCTGCGTAGATTTCTTCAACTGGTGTCACCTAGTTCCTCAGCATGGTGTCTGCAACC
ACAAGTTTTACGGAAAACAATGCTGCAAGTCATGCACAAGGAAGATCTGA

>SGPR_060_SEQ ID NO:19
ATGGACGGCCGCGGGGCTTTCTGGACAGTGGCCATTCCCAGAGCCAGGCAGGAAGGCCTCGGGAG
GCTGGGGCTCCCGTTCCCGGTGAAGCGGACGCCGCCAGCGCCCCAGAACCCAGGAGGAAGCACAC
AGGCCCCACAGAGAGTGGTTGGCAAGAGTCACTCGGGGATTAGGATGCCGGCCAAATCGCGGAAT
TTGAGGCTGGAATCCAAGCTCAACAGGAAAGTAGTGAAATACAAATGGGGAAAACAGGGCTCTG
GAGCGGGGAGGGAGCTGGTGCCGGCATTTCCCACCAACGCCGGTTTAGGAAGACGGGACCGATGC
CGGCCGCCCCCTGCTGGAGGGGATGTGGCATCTCACGGGCTGCCAGGGAGCGGGGTTGGCTACTC
CTGCAACCAGCGTGAAGAGGGTCTCAGGGGAGGCTGTGGTGGGATCCCCCACGTGCCCTTGTTCCT
CTCACCGTTACCTCTGGATGCCTCGGGGCAAAGGCCTTCTTCCACCTATAGACAGAGTCTACGCAG
GGGTCTTGGAACCCGGGCACACCAGTCCCCAGCTAACGAAATCCCCGAGTTGGGGGATTTGAGAG
GGTCACGTTTGGCCCAAGAACCCGCAGTCCTCTTTGGTCTTCGGCCCTCTATTTCTAAGCGTGGGCT
TCTGGCACGGCGGCTCTGGGCACAGCCCATGCTGCTTTCGGGCTGGGTGGTTTCAACGACGACAAC
AATTATCACAGTGACGGTGACCTTCACCCCAACAGGACTGCTGTGTGTGAAGCACTCAAGAGGGC
CCCTACAACCAACCTGCCAGGAGTCGGCTCCTGAAAACAGGGTCGGAAAAGCGCTAATTACTTTTT
CCAAAGGCTGGAGGGCTTCACTCCGGCTGGCGCCGCCGCCTAGCGCGCTCCTGCTTCGCCGCCACG
GTCCGGGGGGGCTGCCGGTCCCGGGTACCATGTGTGACGGCGCCCTGCTGCCTCCGCTCGTCCTGC
CCGTGCTGCTGCTGCTGGTTTGGGGACTGGACCCGGGCACAGGTAGCGCCCCCTCCCACAGCCCTC
TTCACCCCGCGTCCTGCGGCTACCTTCCCTCTGCGTTCTCGCGGCGTCCTGGCGGCCCGGGGGCGG
CGGCGGGACCGCTGACGGCGCCCGAGCGGAGGAGGCGCGGGCCGCGGCCGGAGTACGGGAATCG
GGTGGCTCCGTGGCAGGCGCGCCGCCGCCGGGTCTCCGCTCGCCGATGCGCGGCGCCGTTCCGGG
AGGTGCTCGCGCGGCTGCGCCGGAGACCCTCCCCGGGTGGCGCGGGCCAGCGTGGAGCTGTCGGC
GACGCGGCGGCCGACGTGGAGGTGGTGCTCCCGTGGCGGGTGCGCCCCGACGACGTGCACCTGCC
GCCGCTGCCCGCAGCCCCCGGGCCCCGACGGCGGCGACGCCCCCGCACGCCCCCAGCCGCCCCGC
GCGCCCGGCCCGGAGAGCGCGCCCTGCTGCTGCACCTGCCGGCCTTCGGGCGCGACCTGTACCTTC
AGCTGCGCCGCGACCTGCGCTTCCTGTCCCGAGGCTTCGAGGTGGAGGAGGCGGGCGCGGCCCGG
CGCCGCGGCCGCCCCGCCGAGCTGTGCTTCTACTCGGGCCGTGTGCTCGGCCACCCCGGCTCCCTC
GTCTCGCTCAGCGCCTGCGGCGCCGCCGGCGGCCTGGTTGGCCTCATTCAGCTTGGGCAGGAGCAG
GTGCTAATCCAGCCCCTCAACAACTCCCAGGGCCCATTCAGTGGACGAGAACATCTGATCAGGCG
CAAATGGTCCTTGACCCCCAGCCCTTCTGCTGAGGCCCAGAGACCTGAGCAGGTCTGCAAGGTTCT
AACAGAAAAGAAGAAGCCGACGTGGGGCAGGCCTTCGCGGGACTGGCGGGAGCGGAGGAACGCT
ATCCGGCTCACCAGCGAGCACACGGTGGAGACCCTGGTGGTGGCCGACGCCGACATGGTGCAGTA

Figure 1K

CCACGGGGCCGAGGCCGCCCAGAGGTTCATCCTGACCGTCATGAACATGGTATACAATATGTTTCA
GCACCAGAGCCTGGGGATTAAAATTAACATTCAAGTGACCAAGCTTGTCCTGCTACGACAACGTCC
CGCTAAGTTGTCCATTGGGCACCATGGTGAGCGGTCCCTGGAGAGCTTCTGTCACTGGCAGAACGA
GGAGTATGGAGGAGCGCGATACCTCGGCAATAACCAGGTTCCCGGCGGGAAGGACGACCCGCCCC
TGGTGGATGCTGCCGTGTTTGTGACCAGGACAGATTTATGTGTACACAAAGATGAACCGTGTGACA
CTGTTGGAATTGCTTACTTAGGAGGTGTGTGCAGTGCTAAGAGGAAGTGTGTGCTTGCCGAAGACA
ATGGTCTCAATTTGGCCTTTACCATCGCCCATGAGCTGGGCCACAACTTGGGCATGAACCACGACG
ATGACCACTCATCTTGCGCTGGCAGGTCCCACATCATGTCAGGAGAGTGGGTGAAAGGCCGGAAC
CCAAGTGACCTCTCTTGGTCCTCCTGCAGCCGAGATGACCTTGAAAACTTCCTCAAGTCAAAAGTC
AGCACCTGCTTGCTAGTCACGGACCCCAGAAGCCAGCACACAGTACGCCTCCCGCACAAGCTGCC
GGGCATGCACTACAGTGCCAACGAGCAGTGCCAGATCCTGTTTGGCATGAATGCCACCTTCTGCAG
AAACATGGAGCATCTAATGTGTGCTGGACTGTGGTGCCTGGTAGAAGGAGACACATCCTGCAAGA
CCAAGCTGGACCCTCCCCTGGATGGCACCGAGTGTGGGGCAGACAAGTGGTGCCGCGCGGGGGAG
TGCGTGAGCAAGACGCCCATCCCGGAGCATGTGGACGGAGACTGGAGCCCGTGGGGCGCCTGGAG
CATGTGCAGCCGAACATGTGGGACGGGAGCCCGCTTCCGGCAGAGGAAATGTGACAACCCCCCCC
CTGGGCCTGGAGGCACACACTGCCCGGGTGCCAGTGTAGAACATGCGGTCTGCGAGAACCTGCCC
TGCCCCAAGGGTCTGCCCAGCTTCCGGGACCAGCAGTGCCAGGCACACGACCGGCTGAGCCCCAA
GAAGAAAGGCCTGCTGACAGCCGTGGTGGTTGACGATAAGCCATGTGAACTCTACTGCTCGCCCCT
CGGGAAGGAGTCCCCACTGCTGGTGGCCGACAGGGTCCTGGACGGTACACCCTGCGGGCCCTACG
AGACTGATCTCTGCGTGCACGGCAAGTGCCAGAAAATCGGCTGTGACGGCATCATCGGGTCTGCA
GCCAAAGAGGACAGATGCGGGGTCTGCAGCGGGGACGGCAAGACCTGCCACTTGGTGAAGGGCG
ACTTCAGCCACGCCCGGGGGACAGGTTATATCGAAGCTGCCGTCATTCCTGCTGGAGCTCGGAGG
ATCCGTGTGGTGGAGGATAAACCTGCCCACAGCTTTCTGGCCGTGGTGGTTGACGATAAGCCATGT
GAACTCTACTGCTCGCCCCTCGGGAAGGAGTCCCCACTGCTGGTGGCCGACAGGGTCCTGGACGGT
ACACCCTGCGGGCCCTACGAGACTGATCTCTGCGTGCACGGCAAGTGCCAGAAAATCGGCTGTGA
CGGCATCATCGGGTCTGCAGCCAAAGAGGACAGATGCGGGGTCTGCAGCGGGGACGGCAAGACCT
GCCACTTGGTGAAGGGTGACTTCAGCCACGCCCGGGGGACAGGTTATATCGAAGCTGCCGTCATTC
CTGCTGGAGCTCGGAGGATCCGTGTGGTGGAGGATAAACCTGCCCACAGCTTTCTAGCTCTCAAAG
ACTCGGGTAAGGGGTCCATCAACAGTGACTGGAAGATAGAGCTCCCCGGAGAGTTCCAGATTGCA
GGCACAACTGTTCGCTATGTGAGAAGGGGGCTGTGGGAGAAGATCTCTGCCAAGGGACCAACCAA
ACTACCGCTGCACTTGATGGTGTTGTTATTTCACGACCAAGATTATGGAATTCATTATGAATACACT
GTTCCTGTAAACCGCACTGCGGAAAATCAAAGCGAACCAGAAAAACCGCAGGACTCTTTGTTCAT
CTGGACCCACAGCGGCTGGGAAGGGTGCAGTGTGCAGTGCGGCGGAGGTGAGTGGCCGTGGTCCA
TGACCTGTTGGGTGTGGGGTTTTGCTGAAGGAAGGAGAAAGGCATCTGTGGCCAGCACGCAGAGT
GTGAGACATCTGCAACCTGTAGCTCCATGGGAATTTAACCATATCCCACCGAAAATCTCTCTGCAG
AATACTTGGACAGAGTCTTCCCAACTCCCACACTAG

>SGPR_068_SEQ ID NO:20
ATGGCTCCACTCCGCGCGCTGCTGTCCTACCTGCTGCCTTTGCACTGTGCGCTCTGCGCCGCCGCGG
GCAGCCGGACCCCAGAGCTGCACCTCTCTGGAAAGCTCAGTGACTATGGTGTGACAGTGCCCTGC
AGCACAGACTTTCGGGGACGCTTCCTCTCCCACGTGGTGTCTGGCCCAGCAGCAGCCTCTGCAGGG
AGCATGGTAGTGGACACGCCACCCACACTACCACGACACTCCAGTCACCTCCGGGTGGCTCGCAG
CCCTCTGCACCCAGGAGGGACCCTGTGGCCTGGCAGGGTGGGGCGCCACTCCCTCTACTTCAATGT
CACTGTTTTCGGGAAGGAACTGCACTTGCGCCTGCGGCCCAATCGGAGGTTGGTAGTGCCAGGATC
CTCAGTGGAGTGGCAGGAGGATTTTCGGGAGCTGTTCCGGCAGCCCTTACGGCAGGAGTGTGTGT
ACACTGGAGGTGTCACTGGAATGCCTGGGGCAGCTGTTGCCATCAGCAACTGTGACGGATTGGCG
GGCCTCATCCGCACAGACAGCACCGACTTCTTCATTGAGCCTCTGGAGCGGGGCCAGCAGGAGAA
GGAGGCCAGCGGGAGGACACATGTGGTGTACCGCCGGGAGGCCGTCCAGCAGGAGTGGGCAGAA
CCTGACGGGGACCTGCACAATGAAGCCTTTGGCCTGGGAGACCTTCCCAACCTGCTGGGCCTGGTG
GGGGACCAGCTGGGCGACACAGAGCGGAAGCGGCGGCATGCCAAGCCAGGCAGCTACAGCATCG
AGGTGCTGCTGGTGGTGGACGACTCGGTGGTTCGCTTCCATGGCAAGGAGCATGTGCAGAACTAT
GTCCTCACCCTCATGAATATCGTAGATGAGATTTACCACGATGAGTCCCTGGGGGTTCATATAAAT
ATTGCCCTCGTCCGCTTGATCATGGTTGGCTACCGACAGTCCCTGAGCCTGATCGAGCGCGGGAAC
CCCTCACGCAGCCTGGAGCAGGTGTGTCGCTGGGCACACTCCCAGCAGCGCCAGGACCCCAGCCA
CGCTGAGCACCATGACCACGTTGTGTTCCTCACCCGGCAGGACTTTGGGCCCTCAGGGTATGCACC
CGTCACTGGCATGTGTCACCCCCTGAGGAGCTGTGCCCTCAACCATGAGGATGGCTTCTCCTCAGC
CTTCGTGATAGCTCATGAGACCGGCCACGTGCTCGGCATGGAGCATGACGGTCAGGGGAATGGCT
GTGCAGATGAGACCAGCCTGGGCAGCGTCATGGCGCCCCTGGTGCAGGCTGCCTTCCACCGCTTCC

Figure 1L

ATTGGTCCCGCTGCAGCAAGCTGGAGCTCAGCCGCTACCTCCCCTCCTACGACTGCCTCCTCGATG
ACCCCTTTGATCCTGCCTGGCCCCAGCCCCCAGAGCTGCCTGGGATCAACTACTCAATGGATGAGC
AGTGCCGCTTTGACTTTGGCAGTGGCTACCAGACCTGCTTGGCATTCAGGACCTTTGAGCCCTGCA
AGCAGCTGTGGTGCAGCCATCCTGACAACCCGTACTTCTGCAAGACCAAGAAGGGGCCCCCGCTG
GATGGGACTGAGTGTGCACCCGGCAAGTGGTGCTTCAAAGGTCACTGCATCTGGAAGTCGCCGGA
GCAGACATATGGCCAGGATGGAGGCTGGAGCTCCTGGACCAAGTTTGGGTCATGTTCGCGGTCAT
GTGGGGGCGGGGTGCGATCCCGCAGCCGGAGCTGCAACAACCCCTCCCCAGCCTATGGAGGCCGC
CCGTGCTTAGGGCCCATGTTCGAGTACCAGGTCTGCAACAGCGAGGAGTGCCCTGGGACCTACGA
GGACTTCCGGGCCCAGCAGTGTGCCAAGCGCAACTCGTACTATGTGCACCAGAATGCCAAGCACA
GCTGGGTGCCCTACGAGCCTGACGATGACGCCCAGAAGTGTGAGCTGATCTGCCAGTCGGCGGAC
ACGGGGGACGTGGTGTTCATGAACCAGGTGGTTCACGATGGGACACGCTGCAGCTACCGGGACCC
ATACAGCGTCTGTGCGCGTGGCGAGTGTGTGCCTGTCGGCTGTGACAAGGAGGTGGGGTCCATGA
AGGCGGATGACAAGTGTGGAGTCTGCGGGGGTGACAACTCCCACTGCAGGACTGTGAAGGGGACG
CTGGGCAAGGCCTCCAAGCAGGCAGGAGCTCTCAAGCTGGTGCAGATCCCAGCAGGTGCCAGGCA
CATCCAGATTGAGGCACTGGAGAAGTCCCCCCACCGCATTGTGGTGAAGAACCAGGTCACCGGCA
GCTTCATCCTCAACCCCAAGGGCAAGGAAGCCACAAGCCGGACCTTCACCGCCATGGGCCTGGAG
TGGGAGGATGCGGTGGAGGATGCCAAGGAAAGCCTCAAGACCAGCGGGCCCCTGCCTGAAGCCAT
TGCCATCCTGGCTCTCCCCCCAACTGAGGGTGGCCCCCGCAGCAGCCTGGCCTACAAGTACGTCAT
CCATGAGGACCTGCTGCCCCTTATCGGGAGCAACAATGTGCTCCTGGAGGAGATGGACACCTATG
AGTGGGCGCTCAAGAGCTGGGCCCCCTGCAGCAAGGCCTGTGGAGGAGGGATCCAGTTCACCAAA
TACGGCTGCCGGCGCAGACGAGACCACCACATGGTGCAGCGACACCTGTGTGACCACAAGAAGAG
GCCCAAGCCCATCCGCCGGCGCTGCAACCAGCACCCGTGCTCTCAGCCTGTGTGGGTGACGGAGG
AGTGGGGTGCCTGCAGCCGGAGCTGTGGGAAGCTGGGGGTGCAGACACGGGGGATACAGTGCCTG
ATGCCCCTCTCCAATGGAACCCACAAGGTCATGCCGGCCAAAGCCTGTGCCGGGGACCGGCCTGA
GGCCCGACGGCCCTGTCTCCGAGTGCCCTGCCCAGCCCAGTGGAGGCTGGGAGCCTGGTCCCAGT
GCTCTGCCACCTGTGGAGAGGGCATCCAGCAGCGGCAGGTGGTGTGCAGGACCAACGCCAACAGC
CTCGGGCATTGCGAGGGGGATAGGCCAGACACTGTCCAGGTCTGCAGCCTGCCTGCCTGTGGAGC
GGAGCCCTGCACGGGAGACAGGTCTGTCTTCTGCCAGATGGAAGTGCTCGATCGCTACTGCTCCAT
TCCCGGCTACCACCGGCTCTGCTGTGTGTCCTGCATCAAGAAGGCCTCGGGCCCCAACCCTGGCCC
AGACCCTGGCCCAACCTCACTGCCCCCCTTCTCCACTCCTGGAAGCCCCTTACCAGGACCCCAGGA
CCCTGCAGATGCTGCAGAGCCTCCTGGAAAGCCAACGGGATCAGAGGACCATCAGCATGGCCGAG
CCACACAGCTCCCAGGAGCTCTGGATACAAGCTCCCCAGGGACCCAGCATCCCTTTGCCCCTGAGA
CACCAATCCCTGGAGCATCCTGGAGCATCTCCCCTACCACCCCCGGGGGGCTGCCTTGGGGCTGGA
CTCAGACACCTACGCCAGTCCCTGAGGACAAAGGGCAACCTGGAGAAGACCTGAGACATCCCGGC
ACCAGCCTCCCTGCTGCCTCCCCGGTGACATGA

>SGPR_096_SEQ ID NO:21
ATGCAGTTTGTATCCTGGGCCACACTGCTAACGCTCCTGGTGCGGGACCTGGCCGAGATGGGGAGC
CCAGACGCCGCGGCGGCCGTGCGCAAGGACAGGCTGCACCCGAGGCAAGTGAAATTATTAGAGAC
CCTGAGCGAATACGAAATCGTGTCTCCCATCCGAGTGAACGCTCTCGGAGAACCCTTTCCCACGAA
CGTCCACTTCAAAAGAACGCGACGGAGCATTAACTCTGCCACTGACCCCTGGCCTGCCTTCGCCTC
CTCCTCTTCCTCCTCTACCTCCTCCCAGGCGCATTACCGCCTCTCTGCCTTCGGCCAGCAGTTTCTAT
TTAATCTCACCGCCAATGCCGGATTTATCGCTCCACTGTTCACTGTCACCCTCCTCGGGACGCCCGG
GGTGAATCAGACCAAGTTTTATTCCGAAGAGGAAGCGGAACTCAAGCACTGTTTCTACAAAGGCT
ATGTCAATACCAACTCCGAGCACACGGCCGTCATCAGCCTCTGCTCAGGAATGCTGGGCACATTCC
GGTCTCATGATGGGGATTATTTTATTGAACCACTACAGTCTATGGATGAACAAGAAGATGAAGAG
GAACAAAACAAACCCCACATCATTTATAGGCGCAGCGCCCCCCAGAGAGAGCCCTCAACAGGAAG
GCATGCATGTGACACCTCAGAACACAAAAATAGGCACAGTAAAGACAAGAAGAAAACCAGAGCA
AGAAAATGGGGAGAAAGGATTAACCTGGCTGGTGACGTAGCAGCATTAAACAGCGGCTTAGCAAC
AGAGGCATTTTCTGCTTATGGTAATAAGACGGACAACACAAGAGAAAAGAGGACCCACAGAAGG
ACAAAACGTTTTTTATCCTATCCACGGTTTGTAGAAGTCTTGGTGGTGGCAGACAACAGAATGGTT
TCATACCATGGAGAAAACCTTCAACACTATATTTTAACTTTAATGTCAATTGTAGCCTCTATCTATA
AAGACCCAAGTATTGGAAATTTAATTAATATTGTTATTGTGAACTTAATTGTGATTCATAATGAAC
AGGATGGGCCTTCCATATCTTTTAATGCTCAGACAACATTAAAAAACTTTTGCCAGTGGCAGCATT
CGAAGAACAGTCCAGGTGGAATCCATCATGATACTGCTGTTCTCTTAACAAGACAGGATATCTGCA
GAGCTCACGACAAATGTGATACCTTAGGCCTGGCTGAACTGGGAACCATTTGTGATCCCTATAGAA
GCTGTTCTATTAGTGAAGATAGTGGATTGAGTACAGCTTTTACGATCGCCCATGAGCTGGGCCATG
TGTTTAACATGCCTCATGATGACAACAACAAATGTAAAGAAGAAGGAGTTAAGAGTCCCCAGCAT

Figure 1M

GTCATGGCTCCAACACTGAACTTCTACACCAACCCCTGGATGTGGTCAAAGTGTAGTCGAAAATAT
ATCACTGAGTTTTTAGACACTGGTTATGGCGAGTGTTTGCTTAACGAACCTGAATCCAGACCCTAC
CCTTTGCCTGTCCAACTGCCAGGCATCCTTTACAACGTGAATAAACAATGTGAATTGATTTTTGGA
CCAGGTTCTCAGGTGTGCCCATATATGATGCAGTGCAGACGGCTCTGGTGCAATAACGTCAATGGA
GTACACAAAGGCTGCCGGACTCAGCACACACCCTGGGCCGATGGGACGGAGTGCGAGCCTGGAAA
GCACTGCAAGTATGGATTTTGTGTTCCCAAAGAAATGGATGTCCCCGTGACAGATGGATCCTGGGG
AAGTTGGAGTCCCTTTGGAACCTGCTCCAGAACATGTGGAGGGGGCATCAAAACAGCCATTCGAG
AGTGCAACAGACCAGAACCAAAAAATGGTGGAAAATACTGTGTAGGACGTAGAATGAAATTTAA
GTCCTGCAACACGGAGCCATGTCTCAAGCAGAAGCGAGACTTCCGAGATGAACAGTGTGCTCACT
TTGACGGGAAGCATTTTAACATCAACGGTCTGCTTCCCAATGTGCGCTGGGTCCCTAAATACAGTG
GAATTCTGATGAAGGACCGGTGCAAGTTGTTCTGCAGAGTGGCAGGGAACACAGCCTACTATCAG
CTTCGAGACAGAGTGATAGATGGAACTCCTTGTGGCCAGGACACAAATGATATCTGTGTCCAGGG
CCTTTGCCGGCAAGCTGGATGCGATCATGTTTTAAACTCAAAAGCCCGGAGAGATAAATGTGGGG
TTTGTGGTGGCGATAATTCTTCATGCAAAACAGTGGCAGGAACATTTAATACAGTACATTATGGTT
ACAATACTGTGGTCCGAATTCCAGCTGGTGCTACCAATATTGATGTGCGGCAGCACAGTTTCTCAG
GGGAAACAGACGATGACAACTACTTAGCTTTATCAAGCAGTAAAGGTGAATTCTTGCTAAATGGA
AACTTTGTTGTCACAATGGCCAAAAGGGAAATTCGCATTGGGAATGCTGTGGTAGAGTACAGTGG
GTCCGAGACTGCCGTAGAAAGAATTAACTCAACAGATCGCATTGAGCAAGAACTTTTGCTTCAGGT
TTTGTCGGTGGGAAAGTTGTACAACCCCGATGTACGCTATTCTTTCAATATTCCAATTGAAGATAA
ACCTCAGCAGTTTTACTGGAACAGTCATGGGCCATGGCAAGCATGCAGTAAACCCTGCCAAGGGG
AACGGAAACGAAAACTTGTTTGCACCAGGGAATCTGATCAGCTTACTGTTTCTGATCAAAGATGCG
ATCGGCTGCCCCAGCCTGGACACATTACTGAACCCTGTGGTACAGACTGTGACCTGAGGTGGCATG
TTGCCAGCAGGAGTGAATGTAGTGCCCAGTGTGGCTTGGGTTACCGCACATTGGACATCTACTGTG
CCAAATATAGCAGGCTGGATGGGAAGACTGAGAAGGTTGATGATGGTTTTTGCAGCAGCCATCCC
AAACCAAGCAACCGTGAAAAATGCTCAGGGGAATGTAACACGGGTGGCTGGCGCTATTCTGCCTG
GACTGAATGTTCAAAAAGCTGTGACGGTGGGACCCAGAGGAGAAGGGCTATTTGTGTCAATACCC
GAAATGATGTACTGGATGACAGCAAATGCACACATCAAGAGAAAGTTACCATTCAGAGGTGCAGT
GAGTTCCCTTGTCCACAGTGGAAATCTGGAGACTGGTCAGAGTGCTTGGTCACCTGTGGAAAAGG
GCATAAGCACCGCCAGGTCTGGTGTCAGTTTGGTGAAGATCGATTAAATGATAGAATGTGTGACCC
TGAGACCAAGCCAACATCTATGCAGACTTGTCAGCAGCCGGAATGTGCATCCTGGCAGGCGGGTC
CCTGGGGACAGTGCAGTGTCACTTGTGGACAGGGATACCAGCTAAGAGCAGTGAAATGCATCATT
GGGACTTATATGTCAGTGGTAGATGACAATGACTGTAATGCAGCAACTAGACCAACTGATACCCA
GGACTGTGAATTACCATCATGTCATCCTCCCCCAGCTGCCCCGGAAACGAGGAGAAGCACATACA
GTGCACCAAGAACCCAGTGGCGATTTGGGTCTTGGACCCCATGCTCAGCCACTTGTGGGAAAGGT
ACCCGGATGAGATACGTCAGCTGCCGAGATGAGAATGGCTCTGTGGCTGACGAGAGTGCCTGTGC
TACCCTGCCTAGACCAGTGGCAAAGGAAGAATGTTCTGTGACACCCTGTGGGCAATGGAAGGCCT
TGGACTGGAGCTCTTGCTCTGTGACCTGTGGGCAAGGTAGGGCAACCCGGCAAGTGATGTGTGTCA
ACTACAGTGACCACGTGATCGATCGGAGTGAGTGTGACCAGGATTATATCCCAGAAACTGACCAG
GACTGTTCCATGTCACCATGCCCTCAAAGGACCCCAGACAGTGGCTTAGCTCAGCACCCCTTCCAA
AATGAGGACTATCGTCCCCGGAGCGCCAGCCCCAGCCGCACCCATGTGCTCGGTGGAAACCAGTG
GAGAACTGGCCCCTGGGGAGCATGTTCCAGTACCTGTGCTGGCGGATCCCAGCGGCGTGTTGTTGT
ATGTCAGGATGAAAATGGATACACCGCAAACGACTGTGTGGAGAGAATAAAACCTGATGAGCAA
AGAGCCTGTGAATCCGGCCCTTGTCCTCAGTGGGCTTATGGCAACTGGGGAGAGTGCACTAAGCTG
TGTGGTGGAGGCATAAGAACAAGACTGGTGGTCTGTCAGCGGTCCAACGGTGAACGGTTTCCAGA
TTTGAGCTGTGAAATTCTTGATAAACCTCCCGATCGTGAGCAGTGTAACACACATGCTTGTCCACA
CGACGCTGCATGGAGTACTGGCCCTTGGAGCTCGTGTTCTGTCTCTTGTGGTCGAGGGCATAAACA
ACGAAATGTTTACTGCATGGCAAAAGATGGAAGCCATTTAGAAAGTGATTACTGTAAGCACCTGG
CTAAGCCACATGGGCACAGAAAGTGCCGAGGAGGAAGATGCCCCAAATGGAAAGCTGGCGCTTG
GAGTCAGTGCTCTGTGTCCTGTGGCCGAGGCGTACAGCAGAGGCATGTGGGCTGTCAGATCGGAA
CACACAAAATAGCCAGAGAGACCGAGTGCAACCCATACACCAGACCGGAGTCGGAACGCGACTG
CCAAGGCCCACGGTGTCCCCTCTACACTTGGAGGGCAGAGGAATGGCAAGAATGCACCAAGACCT
GCGGCGAAGGCTCCAGGTACCGCAAGGTGGTGTGTGTGGATGACAACAAAAACGAGGTGCATGG
GGCACGCTGTGACGTGAGCAAGCGGCCGGTGGACCGTGAAAGCTGTAGTTTGCAACCCTGCGAGT
ATGTCTGGATCACAGGAGAATGGTCAGAGTGCTCAGTGACCTGTGGAAAAGGCTACAAACAAAGG
CTTGTCTCGTGCAGCGAGATTTACACCGGGAAGGAGAATTATGAATACAGCTACCAAACCACCAT
CAACTGCCCAGGCACGCAGCCCCCCAGTGTTCACCCCTGTTACCTGAGGGACTGCCCTGTCTCGGC
CACCTGGAGAGTTGGCAACTGGGGGAGCTGCTCAGTGTCTTGTGGTGTTGGAGTGATGCAGAGAT
CTGTGCAATGTTTAACCAATGAGGACCAACCCAGCCACTTATGCCACACTGATCTGAAGCCAGAA

Figure 1N

GAACGAAAAACCTGCCGTAATGTCTATAACTGTGAGTTACCCCAGAATTGCAAGGAGGTAAAAAG
ACTTAAAGGTGCCAGTGAAGATGGTGAATATTTCCTGATGATTAGAGGAAAGCTTCTGAAGATATT
CTGTGCGGGGATGCACTCTGACCACCCCAAAGAGTACGTGACACTGGTGCATGGAGACTCTGAGA
ATTTCTCCGAGGTTTATGGGCACAGGTTACACAACCCAACAGAATGTCCCTATAACGGGAGCCGGC
GCGATGACTGCCAATGTCGGAAGGATTACACGGCCGCTGGGTTTTCCAGTTTTCAGAAAATCAGAA
TAGACCTGACCAGCATGCAGATAATCACCACTGACTTACAGTTTGCAAGGACAAGCGAAGGACAT
CCCGTCCCTTTTGCCACAGCCGGGGATTGCTACAGCGCTGCCAAGTGCCCACAGGGTCGTTTTAGC
ATCAACCTTTATGGAACCGGCTTGTCTTTAACTGAATCTGCCAGATGGATATCACAAGGGAATTAT
GCTGTCTCTGACATCAAGAAGTCGCCGGATGGTACCCGAGTCGTAGGGAAATGCGGTGGTTACTGT
GGAAAATGCACTCCATCCTCTGGTACTGGCCTGGAGGTGCGAGTTTTATAG

>SGPR_119_SEQ ID NO:22
ATGTGGGTGGCCAAGTGGCTGACTGGGCTGCTCTACCATCTCTCGCTCTTCATCACCAGGTCTTGG
GAAGTTGACTTCCACCCCAGGCAAGAAGCCCTGGTGAGGACACTGACCTCCTACGAAGTAGTGAT
CCCCGAGCGGGTCAATGAGTTTGGAGAAGTGTTCCCTCAGAGCCACCACTTCAGCCGGCAGAAAC
GCAGCTCCGAGGCGCTGGAACCCATGCCGTTCCGAACCCACTATCGCTTCACTGCCTACGGGCAGC
TCTTCCAGCTGAACCTGACCGCCGATGCATCCTTTCTGGCCGCCGGCTACACCGAGGTGCACTTGG
GAACCCCGGAGCGCGGGGCCTGGGAGAGCGACGCAGGGCCCTCGGACCTGCGCCACTGCTTCTAC
CGCGGCCAGGTCAACTCACAGGAGGATTACAAGGCCGTCGTCAGCTTATGCGGAGGCCTGACGGG
AACATTTAAAGGACAGAACGGTGAATATTTCTTAGAACCTATAATGAAGGCAGATGGGAATGAAT
ATGAAGATGGTCACAACAAGCCACATCTTATATACAGACAAGACTTAAATAACTCTTTTCTGCAGA
CTCTGAAGTATTGCAGTGTGTCAGAAAGTCAAATAAAGGAAACCAGTTTACCCTTTCATACCTACA
GCAACATGAATGAAGATCTTAATGTAATGAAAGAAAGAGTTTTAGGACACACATCAAAAAATGTA
CCATTGAAAGATGAAAGAAGACATTCCAGGAAAAAACGTCTTATATCATATCCAAGATACATTGA
AATTATGGTTACAGCTGATGCTAAAGTGGTTTCTGCTCATGGATCGAATTTGCAAAACTATATACT
GACTCTAATGTCAATTGTTGCAACAATCTACAAAGATCCAAGTATTGGAAATTTGATACACATAGT
AGTGGTAAAATTAGTTATGATTCACCGTGAGGAGGAAGGACCAGTCATTAATTTTGATGGTGCTAC
CACATTAAAGAACTTTTGTTCATGGCAACAAACTCAGAATGACCTTGATGATGTTCACCCTTCCCA
CCATGACACTGCTGTTCTTATCACTAGGGAAGACATTTGTTCATCTAAAGAGAAATGTAACATGTT
AGGTTTATCATATTTAGGTACCATATGTGATCCTTTACAAAGCTGCTTTATTAATGAAGAAAAAGG
ACTCATTTCTGCTTTTACTATAGCCCATGAGCTTGGGCACACACTTGGTGTTCAACATGATGATAAT
CCTAGATGTAAAGAAATGAAAGTTACAAAGTATCATGTAATGGCCCCTGCTTTAAGTTTTCACATG
AGTCCTTGGAGCTGGTCAAACTGTAGTCGGAAATATGTTACTGAATTCCTAGATACTGGTTACGGG
GAATGTCTTCTTGACAAACCAGATGAAGAAATATATAATCTGCCTTCAGAACTTCCTGGATCACGA
TATGATGGAAACAAGCAGTGTGAGCTTGCGTTTGGTCCTGGGTCACAAATGTGTCCCCATATAGAG
AATATATGCATGCATCTGTGGTGCACAAGCACAGAAAAGCTTCACAAAGGCTGTTTCACTCAACAC
GTGCCACCAGCAGATGGAACAGACTGCGGTCCTGGAATGCATTGCCGTCATGGGCTATGTGTAAA
CAAAGAAACGGAAACACGTCCTGTAAATGGTGAATGGGGACCATGGGAACCTTACAGTTCTTGTT
CAAGAACATGTGGAGGCGGAATCGAAAGTGCAACCAGGCGCTGTAATCGTCCTGAGCCAAGAAAC
GGAGGAAATTACTGTGTGGGCCGCAGGATGAAATTTCGATCATGTAATACTGATTCATGTCCAAAA
GGCACACAAGACTTTCGAGAGAAGCAGTGCTCTGATTTTAATGGTAAACATTTGGACATCAGTGGC
ATTCCCTCTAATGTGAGGTGGCTTCCAAGATACAGTGGCATTGGCACAAAGGATCGTTGTAAACTC
TATTGTCAGGTTGCTGGAACCAATTATTTCTACCTATTGAAGGATATGGTTGAAGATGGTACTCCTT
GTGGAACTGAAACTCATGACATCTGTGTTCAAGGCCAGTGTATGGCAGCTGGTTGTGATCACGTGT
TAAACTCCAGTGCCAAGATAGACAAATGTGGAGTGTGTGGTGGGGACAACTCTTCATGCAAGACA
ATAACAGGTGTCTTCAACAGTTCTCATTATGGTTATAATGTTGTTGTAAAGATTCCCGCAGGAGCA
ACAAACGTTGACATTCGTCAGTACAGCTATTCTGGACAACCAGATGACAGTTACCTTGCATTATCT
GACGCTGAAGGGAATTTTCTTTTCAATGGAAATTTTCTTCTAAGTACGTCAAAAAAAGAAATCAAT
GTGCAAGGAACAAGAACTGTTATTGAATACAGTGGATCAAATAACGCAGTTGAAAGAATTAATAG
TACTAATCGACAAGAGAAAGAACTTATTTTGCAGGTGTTGTGTGTGGGTAATTTATACAACCCTGA
TGTACATTATTCCTTCAATATCCCTTTGGAAGAGAGGAGTGACATGTTCACATGGGACCCCTATGG
ACCATGGGAAGGCTGTACCAAAATGTGTCAAGGTCTTCAGCGAAGAAACATAACTTGCATACATA
AGAGTGATCATAGTGTTGTGTCTGATAAAGAATGTGACCACTTGCCACTTCCATCATTTGTTACTCA
AAGTTGCAATACAGACTGTGAACTAAGGTGGCATGTTATTGGCAAAAGTGAATGTTCATCCCAATG
TGGTCAAGGATATAGAACCTTGGACATCCATTGCATGAAGTATTCCATTCATGAAGGACAGACTGT
TCAAGTTGATGACCACTACTGTGGTGACCAGCTTAAACCTCCTACCCAAGAACTATGCCATGGTAA
CTGTGTCTTCACAAGATGGCATTATTCAGAATGGTCTCAGTGTTCCAGGAGTTGTGGAGGAGGGGA
AAGGTCTCGAGAATCTTATTGTATGAATAACTTTGGCCATCGTCTTGCTGACAATGAATGCCAAGA

Figure 1O

ACTGTCCCGAGTGACGAGAGAGAATTGCAATGAATTTTCCTGTCCCAGTTGGGCTGCTAGTGAATG
GAGCGAGTGCCTTGTTACATGTGGTAAAGGAACAAAGCAGCGGCAGGTATGGTGTCAGCTGAATG
TAGATCACTTGAGTGATGGCTTCTGTAATTCAAGTACCAAACCTGAATCTCTGAGTCCATGTGAAC
TTCATACATGTGCTTCCTGGCAAGTAGGACCATGGGGTCCTTGCACAACCACATGTGGACATGGGT
ATCAGATGCGAGATGTTAAATGTGTCAATGAGCTAGCTAGTGCAGTGTTAGAGGACACAGAATGC
CATGAAGCTAGTCGCCCCAGTGACAGACAGAGCTGTGTACTTACACCTTGCTCATTTATTTCTAAA
CTTGAGACCGCTTTATTACCAACTGTTCTCATAAAAAAGATGGCACAATGGCGACATGGTTCTTGG
ACCCCATGCTCCGTATCTTGTGGAAGAGGTACTCAAGCCCGCTATGTAAGCTGTCGTGATGCTCTT
GATAGAATAGCAGATGAATCATATTGTGCCCACTTACCCCGACCTGCTGAAATATGGGACTGTTTT
ACCCCTTGTGGAGAGTGGCAAGCAGGGGATTGGTCACCCTGTTCAGCTTCCTGTGGCCATGGAAAA
ACAACTCGACAAGTTTTATGCATGAACTACCATCAGCCAATTGATGAGAATTACTGTGATCCTGAA
GTTCGCCCTTTGATGGAACAGGAATGTAGCCTGGCAGCCTGCCCTCCTGCACACAGCCACTTTCCT
AGTTCCCCTGTGCAGCCAAGCTATTATCTAAGCACGAATTTGCCATTAACTCAAAAACTTGAAGAT
AATGAAAATCAGGTGGTCCATCCATCAGTCAGAGGAAACCAGTGGAGAACCGGACCATGGGGATC
ATGCTCCAGCAGTTGTTCTGGAGGTCTTCAGCATAGGGCTGTGGTCTGCCAGGATGAAAATGGACA
AAGTGCTAGTTACTGCGATGCAGCCTCCAAGCCTCCAGAGTTACAGCAATGTGGTCCAGGGCCTTG
TCCACAGTGGAACTACGGAAATTGGGGAGAATGTTCACAAACATGTGGAGGAGGAATAAAATCAA
GACTTGTAATATGTCAATTTCCCAATGGCCAAATATTAGAAGATCACAACTGTGAAATTGTAAACA
AGCCACCTAGCGTAATACAGTGTCATATGCATGCTTGCCCTGCTGATGTGTCATGGCATCAGGAAC
CATGGACATCGGAGGATCTTAAAGTGAAATTGCTGCCTCAAAGGACCATCATCTTGTGGGAACTA
ATGAAAAACATATTTTGCCATGGAAAGCACTCACATATGTATTTAATAAATGTCGTTACTGACCAT
CTACTATATCCTAGGCACTGTGATCCAGAGACAATTGAAACATATTTCTTATCCCTATGGAGTTTAC
AGTTTACTTGGGGAGATTTGAAATACTATAAGAACTCACTATAA

>SGPR_143_SEQ ID NO:23
ATGGGGCGCCCTGTCCCGGCTTCAGCCCCGCCTCGCCCTCAGCTTCTCAGGACTCTGGACATTCAG
GTGGCGCTGACCGGCCTGGAGGTCCGAAGGCGGCGGCCTGAGGCTGCACCGGGCACGGGTCGGCC
GCAATCCAGCCTGGGCGGAGCCGGAGTTGCGAGCCGCTGCCTAGAGGCCGAGGAGCTCACAGCTA
TGGGCTGGAGGCCCCGGAGAGCTCGGGGGACCCCGTTGCTGCTGCTGCTACTACTGCTGCTGCTCT
GGCCAGTGCCAGGCGCCGGGGTGCTTCAAGGACATATCCCTGGGCAGCCAGTCACCCCGCACTGG
GTCCTGGATGGACAACCCTGGCGCACCGTCAGCCTGGAGGAGCCGGTCTCGAAGCCAGACATGGG
GCTGGTGGCCCTGGAGGCTGAAGGCCAGGAGCTCCTGCTTGAGCTGGAGAAGAACCACAGGCTGC
TGGCCCCAGGATACATAGAAACCCACTACGGCCCAGATGGGCAGCCAGTGGTGCTGGCCCCCAAC
CACACGGATCATTGCCACTACCAAGGGCGAGTAAGGGGCTTCCCCGACTCCTGGGTAGTCCTCTGC
ACCTGCTCTGGGATGAGTGGCCTGATCACCCTCAGCAGGAATGCCAGCTATTATCTGCGTCCCTGG
CCACCCCGGGGCTCCAAGGACTTCTCAACCCACGAGATCTTTCGGATGGAGCAGCTGCTCACCTGG
AAAGGAACCTGTGGCCACAGGGATCCTGGGAACAAAGCGGGCATGACCAGCCTTCCTGGTGGTCC
CCAGAGCAGGGTCAGGCGAGAAGCGCGCAGGACCCGGAAGTACCTGGAACTGTACATTGTGGCA
GACCACACCCTGTTCTTGACTCGGCACCGAAACTTGAACCACACCAAACAGCGTCTCCTGGAAGTC
GCCAACTACGTGGACCAGCTTCTCAGGACTCTGGACATTCAGGTGGCGCTGACCGGCCTGGAGGT
GTGGACCGAGCGGGACCGCAGCCGCGTCACGCAGGACGCCAACGCCACGCTCTGGGCCTTCCTGC
AGTGGCGCCGGGGGCTGTGGGCGCAGCGGCCCCACGACTCCGCGCAGCTGCTCACGGGCCGCGCC
TTCCAGGGCGCCACAGTGGGCCTGGCGCCCGTCGAGGGCATGTGCCGCGCCGAGAGCTCGGGAGG
CGTGAGCACGGACCACTCGGAGCTCCCCATCGGCGCCGCAGCCACCATGGCCCATGAGATCGGCC
ACAGCCTCGGCCTCAGCCACGACCCCGACGGCTGCTGCGTGGAGGCTGCGGCCGAGTCCGGAGGC
TGCGTCATGGCTGCGGCCACCGGGGTGGTTTATGAGCACCCGTTTCCGCGCGTGTTCAGCGCCTGC
AGCCGCCGCCAGCTGCGCGCCTTCTTCCGCAAGGGGGGCGGCGCTTGCCTCTCCAATGCCCCGGAC
CCCGGACTCCCGGTGCCGCCGGCGCTCTGCGGGAACGGCTTCGTGGAAGCGGGCGAGGAGTGTGA
CTGCGGCCCTGGCCAGGAGTGCCGCGACCTCTGCTGCTTTGCTCACAACTGCTCGCTGCGCCCGGG
GGCCCAGTGCGCCCACGGGGACTGCTGCGTGCGCTGCCTGCTGAAGCCGGCTGGAGCGCTGTGCC
GCCAGGCCATGGGTGACTGTGACCTCCCTGAGTTTTGCACGGGCACCTCCTCCCACTGTCCCCCAG
ACGTTTACCTACTGGACGGCTCACCCTGTGCCAGGGGCAGTGGCTACTGCTGGGATGGCGCATGTC
CCACGCTGGAGCAGCAGTGCCAGCAGCTCTGGGGGCCTGGCTCCCACCCAGCTCCCGAGGCCTGTT
TCCAGGTGGTGAACTCTGCGGGAGATGCTCATGGAAACTGCGGCCAGGACAGCGAGGGCCACTTC
CTGCCCTGTGCAGGGAGGGATGCCCTGTGTGGGAAGCTGCAGTGCCAGGGTGGAAAGCCCAGCCT
GCTCGCACCGCACATGGTGCCAGTGGACTCTACCGTTCACCTAGATGGCCAGGAAGTGACTTGTCG
GGGAGCCTTGGCACTCCCCAGTGCCCAGCTGGACCTGCTTGGCCTGGGCCTGGTAGAGCCAGGCA
CCCAGTGTGGACCTAGAATGGTGTGCCAGAGCAGGCGCTGCAGGAAGAATGCCTTCCAGGAGCTT

CAGCGCTGCCTGACTGCCTGCCACAGCCACGGGGTTTGCAATAGCAACCATAACTGCCACTGTGCT
CCAGGCTGGGCTCCACCCTTCTGTGACAAGCCAGGCTTTGGTGGCAGCATGGACAGTGGCCCTGTG
CAGGCTGAAAACCATGACACCTTCCTGCTGGCCATGCTCCTCAGCATCCTGCTGCCTCTGCTCCCA
GGCGCCGGCCTGGCCTGGTGTTGCTACCGACTCCCAGGAGCCCATCTGCAGCGATGCAGCTGGGG
CTGCAGAAGGGACCCTGCGTGCAGTGGCCCCAAAGATGGCCCACACAGGGACCACCCCCTGGGCG
GCGTTCACCCCATGGAGTTGGGCCCCACAGCCACTGGACAGCCCTGGCCCCTGGACCCTGAGAACT
CTCATGAGCCCAGCAGCCACCCTGAGAAGCCTCTGCCAGCAGTCTCGCCTGACCCCCAAGCAGATC
AAGTCCAGATGCCAAGATCCTGCCTCTGGTGA

>SGPR_164_SEQ ID NO:24
CACGGAGACCGCGGCAGCGGCCGGAGAGCCCGGCCCAGCCCCTTCCCACAGCGCGGCGGTGCGCT
GCCCGGCGCCATGCTTCTGCTGGGCATCCTAACCCTGGCTTTCGCCGGGCGAACCGCTGGAGGCTC
TGAGCCAGAGCGGGAGGTAGTCGTTCCCATCCGACTGGACCCGGACATTAACGGCCGCCGCTACT
ACTGGCGGGGTCCCGAGGACTCCGGGGATCAGGGACTCATTTTTCAGATCACAGCATTTCAGGAG
GACTTTTACCTACACCTGACGCCGGATGCTCAGTTCTTGGCTCCCGCCTTCTCCACTGAGCATCTGG
GCGTCCCCCTCCAGGGGCTCACCGGGGGCTCTTCAGACCTGCGACGCTGCTTCTATTCTGGGGACG
TGAACGCCGAGCCGGACTCGTTCGCTGCTGTGAGCCTGTGCGGGGGGCTCCGCGGAGCCTTTGGCT
ACCGAGGCGCCGAGTATGTCATTAGCCCGCTGCCCAATGCTAGCGCGCCGGCGGCGCAGCGCAAC
AGCCAGGGCGCACACCTTCTCCAGCGCCGGGGTGTTCCGGGCGGGCCTTCCGGAGACCCCACCTCT
CGCTGCGGGGTGGCCTCGGGCTGGAACCCCGCCATCCTACGGGCCCTGGACCCTTACAAGCCGCG
GCGGGCGGGCTTCGGGGAGAGTCGTAGCCGGCGCAGGTCTGGGCGCGCCAAGCGTTTCGTGTCTA
TCCCGCGGTACGTGGAGACGCTGGTGGTCGCGGACGAGTCAATGGTCAAGTTCCACGGCGCGGAC
CTGGAACATTATCTGCTGACGCTGCTGGCAACGGCGGCGCGACTCTACCGCCATCCCAGCATCCTC
AACCCCATCAACATCGTTGTGGTCAAGGTGCTGCTTCTTAGAGATCGTGACTCCGGGCCCAAGGTC
ACCGGCAATGCGGCCCTGACGCTGCGCAACTTCTGTGCCTGGCAGAAGAAGCTGAACAAAGTGAG
TGACAAGCACCCCGAGTACTGGGACACTGCCATCCTCTTCACCAGGCAGGACCTGTGTGGAGCCA
CCACCTGTGACACCCTGGGCATGGCTGATGTGGGTACCATGTGTGACCCCAAGAGAAGCTGCTCTG
TCATTGAGGACGATGGGCTTCCATCAGCCTTCACCACTGCCCACGAGCTGGGCCACGTGTTCAACA
TGCCCCATGACAATGTGAAAGTCTGTGAGGAGGTGTTTGGGAAGCTCCGAGCCAACCACATGATG
TCCCCGACCCTCATCCAGATCGACCGTGCCAACCCCTGGTCAGCCTGCAGTGCTGCCATCATCACC
GACTTCCTGGACAGCGGGCACGGTGACTGCCTCCTGGACCAACCCAGCAAGCCCATCTCCCTGCCC
GAGGATCTGCCGGGCGCCAGCTACACCCTGAGCCAGCAGTGCGAGCTGGCTTTTGGCGTGGGCTC
CAAGCCCTGTCCTTACATGCAGTACTGCACCAAGCTGTGGTGCACCGGGAAGGCCAAGGGACAGA
TGGTGTGCCAGACCCGCCACTTCCCCTGGGCCGATGGCACCAGCTGTGGCGAGGGCAAGCTCTGCC
TCAAAGGGGCCTGCGTGGAGAGACACAACCTCAACAAGCACAGGGTGGATGGTTCCTGGGCCAAA
TGGGATCCCTATGGCCCCTGCTCGCGCACATGTGGTGGGGGCGTGCAGCTGGCCAGGAGGCAGTG
CACCAACCCCACCCCTGCCAACGGGGGCAAGTACTGCGAGGGAGTGAGGGTGAAATACCGATCCT
GCAATCTGGAGCCCTGCCCCAGCTCAGCCTCCGGAAAGAGCTTCCGGGAGGAGCAGTGTGAGGCT
TTCAACGGCTACAACCACAGCACCAACCGGCTCACTCTCGCCGTGGCATGGGTGCCCAAGTACTCC
GGCGTGTCTCCCCGGGACAAGTGCAAGCTCATCTGCCGAGCCAATGGCACTGGCTACTTCTATGTG
CTGGCACCCAAGGTGGTGGTGGACGGCACGCTGTGCTCTCCTGACTCCACCTCCGTCTGTGTCCAA
GGCAAGTGCATCAAGGCTGGCTGTGATGGGAACCTGGGCTCCAAGAAGAGATTCGACAAGTGTGG
GGTGTGTGGGGGAGACAATAAGAGCTGCAAGAAGGTGACTGGACTCTTCACCAAGCCCATGCATG
GCTACAATTTCGTGGTGGCCATCCCCGCAGGCGCCTCAAGCATCGACATCCGCCAGCGCGGTTACA
AAGGGCTGATCGGGGATGACAACTACCTGGCTCTGAAGAACAGCCAAGGCAAGTACCTGCTCAAC
GGGCATTTCGTGGTGTCGGCGGTGGAGCGGGACCTGGTGGTGAAGGGCAGTCTGCTGCGGTACAG
CGGCACGGGCACAGCGGTGGAGAGCCTGCAGGCTTCCCGGCCCATCCTGGAGCCGCTGACCGTGG
AGGTCCTCTCCGTGGGGAAGATGACACCGCCCCGGGTCCGCTACTCCTTCTATCTGCCCAAAGAGC
CTCGGGAGGACAAGTCCTCTCATCCCCCGCACCCCCGGGGAGGAGGACCCTCTGTCTTGCACAACA
GCGTCCTCAGCCTCTCCAACCAGGTGGAGCAGCCGGACGACAGGCCCCCTGCACGCTGGGTGGCT
GGCAGCTGGGGGCCGTGCTCCGCGAGCTGCGGCAGTGGCCTGCAGAAGCGGGCGGTGGACTGCCG
GGGCTCCGCCGGGCAGCGCACGGTCCCTGCCTGTGATGCAGCCCATCGGCCCGTGGAGACACAAG
CCTGCGGGGAGCCCTGCCCCACCTGGGAGCTCAGCGCCTGGTCACCCTGCTCCAAGAGCTGCGGCC
GGGGATTTCAGAGGCGCTCACTCAAGTGTGTGGGCCACGGAGGCCGGCTGCTGGCCCGGGACCAG
TGCAACTTGCACCGCAAGCCCCAGGAGCTGGACTTCTGCGTCCTGAGGCCGTGCTGA

>SGPR_281_SEQ ID NO:25

GCGCCTGACTCACATCTGCTGCTGCTGCCTCCTTTACCAGCTGGGGTTCCTGTCGAATGGGATCGTT
TCAGAGCTGCAGTTCGCCCCCGACCGCGAGGAGTGGGAAGTCGTGTTTCCTGCGCTCTGGCGCCGG
GAGCCGGTGGACCCGGCTGGCGGCAGCGGGGGCAGCGCGGACCCGGGCTGGGTGCGCGGCGTTG
GGGGCGGCGGAAGCGCCCGGGCGCAGGCTGCCGGCAGCTCACGCGAGGTGCGCTACTGTGGCTCC
GGTGCCTTTGGAGGAGCCCGTGGAGGGCCGATCAGAGTCCCGGCTCCGGCCCCCGCCGCCGTCGG
AGGGTGAGGAGGACGAGGAGCTTCGAGTCGCAGGAGCTGCCGCGGGGATCCAGCGGGGCTGCCG
CCTTGTCCCCGGGCGCCCCGGCCTCGTGGCAGCCGCCGCCTCCCCCGCAGCCGCCCCCGTCCCCGC
CCCCGGCCCAGCATGCCGAGCCGGATGGCGACGAAGTGTTGCTGCGGATCCCGGCCTTCTCTCGGG
ACCTGTACCTGCTGCTCCGGAGAGACGGCCGCTTCCTGGCGCCGCGCTTCGCAGTGGAACAGCGGC
CAAATCCCGGCCCCGGCCCCACGGGGGCAGCATCCGCCCCGCAACCTCCCGCGCCACCAGACGCA
GGCTGCTTCTACACCGGAGCTGTGCTGCGGCACCCTGGCTCGCTGGCTTCTTTCAGCACCTGTGGA
GGTGGCCTGGTATTTAACCTTTTCCAACACAAGAGTCTGGGTGTGCAGGTCAATCTTCGTGTGATA
AAGCTTATTCTGCTCCATGAAACTCCACCAGAACTATATATTGGGCATCATGGAGAAAAAATGCTA
GAGAGTTTTTGTAAGTGGCAACATGAAGAATTTGGCAAAAAGAATGATATACATTTAGAGATGTC
AACAAACTGGGGGGAAGACATGACTTCAGTGGATGCAGCTATACTTATAACAAGGAAAGATTTCT
GTGTGCACAAAGATGAACCATGTGATACTGTTGGTATAGCTTACTTGAGTGGAATGTGTAGTGAAA
AGAGAAAATGTATTATTGCTGAAGACAATGGCTTGAATCTTGCTTTTACAATTGCTCATGAAATGG
GTCACAACATGGGCATTAACCATGACAATGACCACCCATCGTGTGCTGATGGTCTTCATATCATGT
CTGGTGAATGGATTAAAGGACAGAATCTTGGTGACGTTTCATGGTCTCGATGTAGCAAGGAAGATT
TGGAAAGATTTCTCAGGTCAAAGGCCAGTAACTGCTTGCTACAAACAAATCCGCAGAGTGTCAATT
CTGTGATGGTTCCCTCCAAGCTGCCAGGGATGACATACACTGCTGATGAACAATGCCAGATCCTTT
TTGGGCCATTGGCTTCTTTTTGTCAGGAGATGCAGCATGTTATTTGCACAGGATTATGGTGCAAGG
TAGAAGGTGAGAAAGAATGCAGAACCAAGCTAGACCCACCAATGGATGGAACTGACTGTGACCTT
GGTAAGTGGTGTAAGGCTGGAGAATGTACCAGCAGGACCTCAGCACCTGAACATCTGGCCGGAGA
GTGGAGCCTGTGGAGTCCTTGTAGCCGAACCTGCAGTGCTGGGATCAGCAGTCGAGAGCGCAAAT
GTCCTGGGCTAGATTCTGAAGCAAGGGATTGTAATGGTCCCAGAAAACAATACAGAATATGTGAG
AATCCACCTTGTCCTGCAGGTTTGCCTGGATTCAGAGACTGGCAATGTCAGGCTTATAGTGTTAGA
ACTTCTCCCCCAAAGCATATACTTCAGTGGCAAGCTGTCCTGGATGAAGAAAAACCATGTGCCTTG
TTTTGCTCTCCTGTTGGAAAAGAACAGCCTATTCTTCTATCAGAAAAAGTGATGGATGGAACTTCT
TGTGGCTATCAGGGATTAGATATCTGTGCAAATGGCAGGTGCCAGAAAGTTGGCTGTGATGGTTTA
TTAGGGTCTCTTGCAAGAGAAGATCATTGTGGTGTATGCAATGGCAATGGAAAATCATGCAAGAT
CATTAAAGGGGATTTTAATCACACCAGAGGAGCAGGTTATGTAGAAGTGCTGGTGATACCTGCTG
GAGCAAGAAGAATCAAAGTTGTGGAGGAAAAGCCGGCACATAGCTATTTAGCTCTCCGAGATGCT
GGCAAACAGTCTATTAATAGTGACTGGAAGATTGAACACTCTGGAGCCTTCAATTTGGCTGGAACT
ACCGTTCATTATGTAAGACGAGGCCTCTGGGAGAAGATCTCTGCCAAAGGTCCTACTACAGCACCT
TTACATCTTCTGGTGCTCCTGTTTCAGGATCAGAATTATGGTCTTCACTATGAATACACTATCCCAT
CAGACCCTCTTCCAGAAAACCAGAGCTCTAAAGCACCTGAGCCCCTCTTCATGTGGACACACACAA
GCTGGGAAGATTGCGATGCCACTTGTGGAGGAGGAGAAAGGAAGACAACAGTGTCCTGCACAAA
AATCATGAGCAAAAATATCAGCATTGTGGACAATGAGAAATGCAAATACTTAACCAAGCCAGAGC
CACAGATTCGAAAGTGCAATGAGCAACCATGTCAAACAAGGTGGATGATGACAGAATGGACCCCT
TGTTCACGAACTTGTGGAAAAGGAATGCAGAGCAGACAAGTGGCCTGTACCCAACAACTGAGCAA
TGGAACACTGATTAGAGCCCGAGAGAGGGACTGCATTGGGCCCAAGCCCGCCTCTGCCCAGCGCT
GTGAGGGCCAGGACTGCATGACCGTGTGGGAGGCGGGAGTGTGGTCTGAGTGTTCAGTCAAGTGT
GGCAAAGGCATACGTCATCGGACCGTTAGATGTACCAACCCAAGAAAGAAGTGTGTCCTCTCTAC
CAGACCCAGGGAGGCTGAAGACTGTGAGGATTATTCAAAATGCTATGTGTGGCGAATGGGTGACT
GGTCTAAGTGCTCAATTACCTGTGGCAAAGGAATGCAGTCCCGTGTAATCCAATGCATGCATAAGA
TCACAGGAAGACATGGAAATGAATGTTTTTCCTCAGAAAAACCTGCAGCATACAGGCCATGCCAT
CTTCAACCCTGCAATGAGAAAATTAATGTAAATACCATAACATCACCCAGACTGGCTGCTCTGACT
TTCAAGTGCCTGGGAGATCAGTGGCCAGTGTACTGCCGAGTGATACGTGAAAAGAACCTATGTCA
GGACATGCGGTGGTATCAGCGCTGCTGTGAAACATGCAGGGACTTCTATGCCCAAAAGCTGCAGC
AGAAGAGTTGA

>SGPR_075_SEQ ID NO:26
TATGATTACTGGGGCTCTGATAGCATGATAGTAACAAATAAAGTCATCGAAATTGTTGGCCTTGCA
AATTCAATGTTCACCCAATTTAAAGTTACTATTGTGCTGTCATCATTGGAGTTATGGTCAGATGAA
AATAAGATTTCTACAGTTGGTGAGGCAGATGAATTATTGCAAAAATTTTTAGAATGGAAACAATCT
TATCTTAACCTAAGGCCTCATGATATTGCATATCTACTANNNTACCCCAAGGAGATAACTCTGGAG

Figure 1R

GCATTTGCAGTTATTGTCACCCAGATGCTGGCACTCAGTCTGGGAATATCATATGACGACCCAAAG
AAATGTCAATGTTCAGAATCCACCTGTATAATGAATCCAGAAGTT

>SGPR_292_SEQ ID NO:27
ATGCTCGCCGCCTCCATCTTCCGTCCGACACTGCTGCTCTGCTGGCTGGCTGCTCCCTGGCCCACCC
AGCCCGAGAGTCTCTTCCACAGCCGGGACCGCTCGGACCTGGAGCCGTCCCCACTGCGCCAGGCC
AAGCCCATTGCCGACCTCCACGCTGCTCAGCGGTTCCTGTCCAGATACGGCTGGTCAGGGGTGTGG
GCGGCCTGGGGGCCCAGTCCCGAGGGGCCGCCGGAGACCCCCAAGGGCGCCGCCCTGGCCGAGGC
GGTGCGCAGGTTCCAGCGGGCGAACGCGCTGCCGGCCAGCGGGGAGCTGGACGCGGCCACCCTAG
CGGCCATGAACCGGCCGCGCTGCGGGGTCCCGGACATGCGCCCACCGCCCCCCTCCGCCCCGCCTT
CGCCCCCGGGCCCGCCCCCCAGAGCCCGCTCCAGGCGCTCCCCGCGGGCGCCGCTGTCCTTGTCCC
GGCGGGGTTGGCAGCCCCGGGGCTACCCCGACGGCGGAGCTGCCCAGGCCTTCTCCAAGAGGACG
CTGAGCTGGCGGCTGCTGGGCGAGGCCCTGAGCAGCCAACTGTCCGCGGCCGACCAGCGGCGCAT
TGTGGCGCTGGCCTTCAGGATGTGGAGCGAGGTGACGCCGCTGGACTTCCGCGAGGACCTGGCCG
CCCCCGGGGCCGCGGTCGACATCAAGCTGGGCTTTGGGAGACGGCGGCACCTGGGCTGTCCGCGG
GCCTTCGATGGGAGCGGGCAGGAGTTTGCACACGCCTGGCGCCTAGGTGACATTCACTTTGACGAC
GACGAGCACTTCACACCTCCCACCAGTGACACGGGCATCAGCCTTCTCAAGGTGGCCGTCCATGAA
ATTGGCCATGTCCTGGGCTTGCCTCACACCTACAGGACGGGATCCATAATGCAACCAAATTACATT
CCCCAGGAGCCTGCCTTTGAGTTGGACTGGTCAGACAGGAAAGCAATTCAAAAGCTGTATGGCTC
CTGTGAGGGATCATTTGATACTGCGTTTGACTGGATTCGCAAAGAGAGAAACCAATATGGAGAGG
TGATGGTGAGATTTAGCACATATTTCTTCCGTAACAGCTGGTACTGGCTTTATGAAAATCGAAACA
ATAGGACACGCTATGGGGACCCTATCCAAATCCTCACTGGCTGGCCTGGAATCCCAACACACAAC
ATAGATGCCTTTGTTCACATCTGGACATGGAAAAGAGATGAACGTTATTTTTTTCAAGGAAATCAA
TACTGGAGATATGACAGTGACAAGGATCAGGCCCTCACAGAAGATGAACAAGGAAAAAGCTATCC
CAAATTGATTTCAGAAGGATTTCCTGGCATCCCAAGTCCCCTAGACACGGCGTTTTATGACCGAAG
ACAGAAGTTAATTTACTTCTTCAAGGAGTCCCTTGTATTTGCATTTGATGTCAACAGAAATCGAGT
ACTTAATTCTTATCCAAAGAGGATTACTGAAGTTTTTCCAGCAGTAATACCACAAAATCATCCTTTC
AGAAATATAGATTCCGCTTATTACTCCTATGCATACAACTCCATTTTCTTTTTCAAAGGCAATGCAT
ACTGGAAGGTAGTTAATGACAAGGACAAACAACAGAATTCCTGGCTTCCTGCTAATGGCTTATTTC
CAAAAAAGTTTATTTCAGAGAAGTGGTTTGATGTTTGTGACGTCCATATCTCCACACTGAACATGT
AA

>SGPR_069_SEQ ID NO:28
ATGGTGGAGAGCGCCGGCCGTGCAGGGCAGAAGCGCCCGGGGTTCCTGGAGGGGGGGCTGCTGCT
GCTGCTGCTGCTGGTGACCGCTGCCCTGGTGGCCTTGGGTGTCCTCTACGCCGACCGCAGAGGGAT
CCCAGAGGCCCAAGAGGTGAGCGAGGTCTGCACCACCCCTGGCTGCGTGATAGCAGCCGCCAGGA
TCCTCCAGAACATGGACCCGACCACGGAACCGTGTGACGACTTCTACCAGTTTGCATGCGGAGGCT
GGCTGCGGCGCCACGTGATCCCTGAGACCAACTCAAGATACAGCATCTTTGACGTCCTCCGCGACG
AGCTGGAGGTCATCCTCAAAGCGGTGCTGGAGAATTCGACTGCCAAGGACCGGCCGGCTGTGGAG
AAGGCCAGGACGCTGTACCGCTCCTGCATGAACCAGAGTGTGATAGAGAAGCGAGGCTCTCAGCC
CCTGCTGGACATCTTGGAGGTGGTGGGAGGCTGGCCGGTGGCGATGGACAGGTGGAACGAGACCG
TAGGACTCGAGTGGGAGCTGGAGCGGCAGCTGGCGCTGATGAACTCACAGTTCAACAGGCGCGTC
CTCATCGACCTCTTCATCTGGAACGACGACCAGAACTCCAGCCGGCACATCATCTACATAGACCAG
CCCACCTTGGGCATGCCCTCCCGAGAGTACTACTTCAACGGCGGCAGCAACCGGAAGGTGCGGGA
AGCCTACCTGCAGTTCATGGTGTCAGTGGCCACGTTGCTGCGGGAGGATGCAAACCTGCCCAGGG
ACAGCTGCCTGGTGCAGGAGGACATGGTGCAGGTGCTGGAGCTGGAGACACAGCTGGCCAAGGCC
ACGGTACCCCAGGAGGAGAGACACGACGTCATCGCCTTGTACCACCGGATGGGACTGGAGGAGCT
GCAAAGCCAATTTGGCCTGAAGGGATTTAACTGGACTCTGTTCATACAAACTGTGCTATCCTCTGT
CAAAATCAAGCTGCTGCCAGATGAGGAAGTGGTGGTCTATGGCATCCCCTACCTGCAGAACCTTG
AAAACATCATCGACACCTACTCAGCCAGGACCATACAGAACTACCTGGTCTGGCGCCTGGTGCTG
GACCGCATTGGTAGCCTAAGCCAGAGATTCAAGGACACACGAGTGAACTACCGCAAGGCGCTGTT
TGGCACAATGGTGGAGGAGGTGCGCTGGCGTGAATGTGTGGGCTACGTCAACAGCAACATGGAGA
ACGCCGTGGGCTCCCTCTACGTCAGGGAGGCGTTCCCTGGAGACAGCAAGAGCATGGTGGAACTC
ATTGACAAGGTGCGGACAGTGTTTGTGGAGACGCTGGACGAGCTGGGCTGGATGGACGAGGAGTC
CAAGAAGAAGGCGCAGGAGAAGGCCATGAGCATCCGGGAGCAGATCGGGCACCCTGACTACATC
CTGGAGGAGATGAACAGGCGCCTGGACGAGGAGTACTCCAATGTGAACTTCTCAGAGGACCTGTA
CTTTGAGAACAGTCTGCAGAACCTCAAGGTGGGCGCCCAGCGGAGCCTCAGGAAGCTTCGGGAAA
AGGTGGACCCAAATCTGATCATCGGGGCGGCGGTGGTCAATGCGTTCTACTCCCCAAACCGAAAC

Figure 1S

CAGATTGTATTCCCTGCCGGGATCCTCCAGCCCCCCTTCTTCAGCAAGGAGCAGCCACAGGCCTTG
AACTTTGGAGGCATTGGGATGGTGATCGGGCACGAGATCACGCACGGCTTTGACGACAATGGTGG
CCGGAACTTCGACAAGAATGGCAACATGATGGATTGGTGGAGTAACTTCTCCACCCAGCACTTCCG
GGAGCAGTCAGAGTGCATGATCTACCAGTACGGCAACTACTCCTGGGACCTGGCAGACGAACAGA
ACGTGAACGGATTCAACACCCTTGGGGAAAACATTGCTGACAACGGAGGGGTGCGGCAAGCCTAT
AAGGCCTACCTCAAGTGGATGGCAGAGGGTGGCAAGGACCAGCAGCTGCCCGGCCTGGATCTCAC
CCATGAGCAGCTCTTCTTCATCAACTATGCCCAGGTGTGGTGCGGGTCCTACCGGCCCGAGTTCGC
CATCCAATCCATCAAGACAGACGTCCACAGTCCCCTGAAGTACAGGGTACTGGGGTCGCTGCAGA
ACCTGGCCGCCTTCGCAGACACGTTCCACTGTGCCCGGGGCACCCCCATGCACCCCAAGGAGCGAT
GCCGCGTGTGGTAG

>SGPR_212_SEQ ID NO:29
ATGAGGCTGAAACTTAAGGGTAGCCATTTGTCAGCAGAAGTAAAGGCCAAGTATTCCCAGAGAGA
AGGCATCGCAGTCAACTGCTGTGACGTGTGTGACGTCCATCTCAAAAGCCTGTGTGAATGTAACTA
CACAGGGTGGCATACGCTGATGTCTGCCCTAGATCCCCACAAGCCTCTAGCTTGGGCCCTCCGTCC
ATTCTCACCTTTCCTCCTCACCTCTAGTCCTGCATTAGAAGCAGCCGGTTCTCCTTCCCAGAGCCCT
CCCTGGCAGATTGTGAACCGACTAGGCCATGCCTCTTCACCTGTGGAGAGTGGCTCTGAAGCAGGG
ACTACAGAAGCATCTCCTACGTTAGGCTGCGTCCAGGAGAGAGGGACTAAGGGATTTCGTTTAGA
AGAAGGGGCAGGGGCTGAGAGTTCGGCTTGTAAATGTGTAGGCGAGAGTGTTGACATACATCACT
TCACACCTGATGAAGGAAAGAGAAGACAGGCTATGAACCTAAGAGGGGTGGAGCGACACCTGCT
GGAACCTGCTGTGGCAGCAGCATCTAGCCAGGGCCGCCAAGTGCTGGGTCGCTCCACCCACAGCA
AGATGGGCCGTGCTGGCCCTCGCAGACTTTTATATTTGCATAAATGGGCCCTGGTGAGGCTTCCAC
ACTGGGACAGAAGAGCAGGCAGGTCCCCGGACAGTGGAGGCTTTTTCTTCATGAATAGTCTTAGA
GCGATTTCGCAGTCATCCACTCGTGGAAGCTTCCTGGCTGGAGTCCGCCCTCCAGTCTCTAGCATCT
TGACAGGAGGGAACCATCTCTGTGGGACCCGCCTCTGCCATGAAATTGCCCATGCCTGGTTTGGCC
TAGCCATCGGGGCCCGAGACTGGACGGAGGAGTGGCTGAGTGAAGGCTTCGCCACTCACTTGGAG
GATGTGTTTTGGGCCACAGCACAGCAGCTGGGTCTTGCTTTTCATACCCTGGCCGTGGACCCAGCA
GTGTGCACATCAGTGTCCCCTGCCACATGGAGTCCTGTGAGGAGAGGCCACATGATAGATACAGA
AAAGGCTCTGGGGTCTGAGTCAGACAGACTTCCAGTCCTGGCTCTGCCATTCGTTGGCTCTGTGAG
TATAGATTCCAGCACAAAGTTTGAAACATTTCCAGAGCAAGTCCGCCAGGCTGACCTCTCCCTTCA
AGTGCGAGACTGGGCTGTTGCTGGCCCTGGCGAGTGCCTGCCTCAGACCGTCCAGGGAGTGGGGG
AGTGCCCTGTGGGGCAGGGGTGGCCACGAGCTGCTTTTTCTCTGCGTTCTCATATGGCCTTTCCTCT
GTGCATGCAGAGAGAAAGACGAGATGCAATGCTCCCCCGAGGAGATGCAGGTGTTAAGAAGCTGC
TCCAGGACCTTCAGCAGGAAGGTGGTATGATCTGCTCAGTATTTGGACGGTGCTGCTCTGCTGCTG
TGTGGAGAGCACCACAGGCAGCCGACGGAAAGCCAGGAGAACGGTTGCAGCCTTGCAGTAGTCCA
TGCAAGAGGCCCTGGAGTGCCTGTGACAGATGCAAGACGCAGACCTATTTGAAATGTGTTTTGGCT
GTGGAACGGGCAGGTCTTTGGTTGATTGAATGTGGAGAAGAGGAAAATGAGTGTATCCAGAATGA
CTTTGAGGTTTTTGAGTTGGACAGTTGGGTAGATGGTGATCCCATTTGTGTGATGATATTTTCTTCT
TATTCATTGGACCCTCAGTTCAGCCTGCGGTTGCTGTTCTTGACTGTGGATGCTGTCAGTCAACCAG
ATGAGGGAGCCGGACTGCATGGGGCTTATGTTCAAGATCACATGGCAGTGGAGAGACTTGGGTCC
AAGCCCTCACCCAGTGGTCATGCTCCTTCCCCTGCAGGTCTCACCTGTGCATCTGGGGCTCAGATG
GGCACTGTTGGCCAGTCTCTACACAAGGGTCAGATTTCACTTCCTCCATTGTTGCAGGGATTGGAC
CTTTCTTCAGGAGGCCCGATTCGAAATCAAATCATTATGTATCAAATATCACTGCCACCTGCCCAC
TCTTTAAATATCCACATTGCCTCTGTGGTTGTGGTGGAAAAGGAAGGCGTGGGGAAGGGAAAGGG
CACGTCAATCTCAGTTGTTGCATTTGGTGCCAAACCCAGTAAAGACAAAACTGGCCACACAAGTG
ACTCGGGAGCATCTGTTATCAAGCATGGACTTAATCCGGAGAAGATCTTCATGCAGGTGCATTATT
TAAAGGGCTACTTCCTTCTTCGGTTTCTTGCCAAAAGACTTGGAGATGAAACCTATTTTTCATTTTT
AAGAAAATTTGTGCACACATTTCATGGACAGCTGATTCTTTCCCAGCCTTCCACAGAACCTTTGCC
CAGCAGCCATCCAGCGAATGTTTGCCACATAGAGAATGTTGCCTGTTTTTCAGTCTTCTCTGGTGA
AGACTTTGGACCTCACTTAATAACATTCCAGGGCTCAACTCCCCAGCCCCCACTCCATGCCACCCC
TAGAGAAGCATCTGAAGCAGCCATGCCTGATGTGTGCGATGAATATGCCTTATCCTCCCGAAACTG
GCTTTCCCAACCAAATAGTTCCTTTCAAAGCACTGAAAGCACCCATGATGCTGTGCCTGGGTCCTT
AGATTTCATCGTGCATGTTGCTGTGGGTGAAGAGGAGCGGTCTCATGTGACTGGGCTCCCTTCCAC
ACTTCAACCCAGGGGAGCGCTGCCCTTTCTGTGA

>SGPR_049_SEQ ID NO:30
ATGGGGCCCCCTTCCAGCTCAGGCTTCTATGTGAGCCGCGCAGTGGCCCTGCTGCTGGCTGGGTTG
GTAGCCGCCCTCCTGCTGGCGCTGGCCGTACTCGCCGCCTTGTACGGCCACTGCGAGCGCGTCCCA

Figure 1T

CCGTCGGAGCTGCCTGGACTCAGGGACTCGGAAGCCGAGTCTTCCCCTCCCCTCAGGCAGAAGCC
GACGCCGACCCCGAAACCCAGCAGTGCACGCGAGCTAGCGGTGACGACCACCCCGAGCAACTGGC
GACCCCCGGGGCCCTGGGACCAGCTACGCCTGCCGCCCTGGCTCGTGCCGCTGCACTACGATCTGG
AGCTGTGGCCGCAGCTGAGGCCCGACGAGCTTCCGGCCGGGTCTTTGCCCTTCACTGGCCGCGTGA
ACATCACGGTGCGCTGCACGGTGGCCACCTCTCGACTGCTGCTGCATAGCCTCTTCCAGGACTGCG
AGCGCGCCGAGGTGCGGGGACCCCTTTCCCCGGGCACTGGGAACGCCACAGTGGGCCGCGTGCCC
GTGGACGACGTGTGGTTCGCGCTGGACACGGAATACATGGTGCTGGAGCTCAGTGAGCCCCTGAA
ACCTGGTAGCAGCTACGAGCTGCAGCTTAGCTTCTCGGGCCTGGTGAAGGAAGACCTCAGGGAGG
GACTCTTCCTCAACGTCTACACCGACCAGGGCGAGCGCAGGGCCCTGTTAGCGTCCCAGCTGGAAC
CAACATTTGCCAGGTATGTTTTCCCTTGTTTTGATGAGCCAGCTCTGAAGGCAACTTTTAATATTAC
AATGATTCATCATCCAAGTTATGTGGCCCTTTCCAACATGCCAAAGCTAGGTCAGTCTGAAAAAGA
AGATGTGAATGGAAGCAAATGGACTGTTACAACCTTTTCCACTACGCCCCACATGCCAACTTACTT
AGTCGCATTTGTTATATGTGACTATGACCACGTCAACAGAACAGAAAGGGGCAAGGAGATACGCA
TCTGGGCCCGGAAAGATGCAATTGCAAATGGAAGTGCAGACTTTGCTTTGAACATCACAGGTCCC
ATCTTCTCTTTTCTGGAGGATTTGTTTAATATCAGTTACTCTCTTCCAAAAACAGATATAATTGCCT
TGCCTAGTTTTGACAACCATGCAATGGAAAACTGGGGACTAATGATATTTGATGAATCAGGATTGT
TGTTGGAACCAAAAGATCAACTGACAGAAAAAAAGACTCTGATCTCCTATGTTGTCTCCCACGAG
ATTGGACACCAGTGGTTTGGAAACTTGGTTACCATGAATTGGTGGAACAATATCTGGCTCAACGAG
GGTTTTGCATCTTATTTTGAGTTTGAAGTAATTAACTACTTTAATCCTAAACTCCCAAGAAATGAGA
TCTTTTTTTCTAACATTTTACATAATATCCTCAGAGAAGATCACGCCCTGGTGACTAGAGCTGTGGC
CATGAAGGTGGAAAATTTCAAAACAAGTGAAATACAGGAACTCTTTGACATATTTACTTACAGCA
AGGGAGCGTCTATGGCCCGGATGCTTTCTTGTTTCTTGAATGAGCATTTATTTGTCAGTGCACTCAA
GTCATATTTGAAGACATTTTCCTACTCAAACGCTGAGCAAGATGATCTATGGAGGCATTTTCAAAT
GGCCATAGATGACCAGAGTACAGTTATTTTGCCAGCAACAATAAAAAACATAATGGACAGTTGGA
CACACCAGAGTGGTTTTCCAGTGATCACTTTAAACGTGTCTACTGGCGTCATGAAACAGGAGCCAT
TTTATCTTGAAAACATTAAAAATCGGACTCTTCTAACCAGCAATGACACATGGATTGTCCCTATTCT
TTGGATAAAAAATGGAACTACACAACCTTTAGTCTGGCTAGATCAAAGCAGCAAAGTATTCCCAG
AAATGCAAGTTTCAGATTCTGACCATGACTGGGTGATTTTGAATTTGAATATGACTGGATATTATA
GAGTTAATTATGATAAATTAGGTTGGAAGAAACTAAATCAACAACTTGAAAAGGATCCTAAGGCT
ATTCCTGTTATTCACAGACTGCAGTTCATTGATGATGCCTTTTCCTTGTCTAAAAACAATTATATTG
AGATTGAAACAGCACTTGAGTTAACCAAGTACCTTGCTGAAGAAGATGAAATTATAGTATGGCAT
ACAGTCTTGGTAAACTTGGTAACCAGGGATCTTGTTTCTGAGGTGAACATCTATGATATATACTCA
TTATTAAAGAGGTACCTATTAAAGAGACTTAATTTAATATGGAATATTTATTCAACTATAATTCGT
GAAAATGTGTTGGCATTACAAGATGACTACTTAGCTCTAATATCACTGGAAAAACTTTTTGTAACT
GCGTGTTGGTTGGGCCTTGAAGACTGCCTTCAGCTGTCAAAAGAACTTTTCGCAAAATGGGTGGAT
CATCCAGAAAATGAAATACCTTATCCAATTAAAGATGTGGTTTTATGTTATGGCATTGCCTTGGGA
AGTGATAAAGAGTGGGACATCTTGTTAAATACTTACACTAATACAACAAACAAAGAAGAAAAGAT
TCAACTTGCTTATGCAATGAGCTGCAGCAAAGACCCATGGATACTTAACAGATATATGGAGTATGC
CATCAGCACATCTCCATTCACTTCTAATGAAACAAATATAATTGAGGTTGTGGCTTCATCTGAAGT
TGGCCGGTATGTCGCAAAAGACTTCTTAGTCAACAACTGGCAAGCTGTGAGTAAAAGGTATGGAA
CACAATCATTGATTAATCTAATATATACAATAGGGAGAACCGTAACTACAGATTTACAGATTGTGG
AGCTGCAGCAGTTTTTCAGTAACATGTTGGAGGAACACCAGAGGATCAGAGTTCATGCCAACTTAC
AGACAATAAAGAATGAAAATCTGAAAAACAAGAAGCTAAGTGCCAGGATAGCTGCGTGGCTAAG
GAGAAACACATAG

>SGPR_026_SEQ ID NO:31
ATGGCGAGCGGCGAGCATTCCCCCGGCAGCGGCGCGGCCCGGCGGCCGCTGCACTCCGCGCAGGC
TGTGGACGTGGCCTCGGCCTCCAACTTCCGGGCCTTTGAGCTGCTGCACTTGCACCTGGACCTGCG
GGCTGAGTTCGGGCCTCCAGGGCCCGGCGCAGGGAGCCGGGGGCTGAGCGGCACCGCGGTCCTGG
ACCTGCGCTGCCTGGAGCCCGAGGGCGCCGCCGAGCTGCGGCTGGACTCGCACCCGTGCCTGGAG
GTGACGGCGGCGGCGCTGCGGCGGGAGCGGCCCGGCTCGGAGGAGCCGCCTGCGGAGCCCGTGA
GCTTCTACACGCAGCCCTTCTCGCACTATGGCCAGGCCCTGTGCGTGTCCTTCCCGCAGCCCTGCCG
CGCCGCCGAGCGCCTCCAGGTGCTGCTCACCTACCGCGTCGGGGAGGGACCCGGGGTTTGCTGGTT
GGCTCCCGAGCAGACAGCAGGAAAGAAGAAGCCCTTCGTGTACACCCAGGGCCAGGCTGTCCTAA
ACCGGGCCTTCTTCCCTTGCTTCGACACGCCTGCTGTTAAATACAAGTATTCAGCTCTTATTGAGGT
CCCAGATGGCTTCACAGCTGTGATGAGTGCTAGCACCTGGGAGAAGAGAGGTCCAAATAAGTTCT
TCTTCCAGATGTGTCAGCCCATCCCCTCCTATCTGATAGCTTTGGCCATCGGAGATCTGGTTTCGGC
TGAAGTTGGACCCAGGAGCCGGGTGTGGGCTGAGCCCTGCCTGATTGATGCTGCCAAGGAGGAGT

Figure 1U

ACAACGGGGTGATAGAAGAATTTTTGGCAACAGGAGAGAAGCTTTTTGGACCTTATGTTTGGGGA
AGGTATGACTTGCTCTTCATGCCACCGTCCTTTCCATTTGGAGGAATGGAGAACCCTTGTCTGACCT
TTGTCACCCCCTGCCTGCTAGCTGGGGACCGCTCCTTGGCAGATGTCATCATCCATGAGATCTCCC
ACAGTTGGTTTGGGAACCTGGTCACCAACGCCAACTGGGGTGAATTCTGGCTCAATGAAGGTTTCA
CCATGTACGCCCAGAGGAGGATCTCCACCATCCTCTTTGGCGCTGCGTACACCTGCTTGGAGGCTG
CAACGGGGCGGGCTCTGCTGCGTCAACACATGGACATCACTGGAGAGGAAAACCCACTCAACAAG
CTCCGCGTGAAGATTGAACCAGGCGTTGACCCGGACGACACCTATAATGAGACCCCCTACGAGAA
AGGTTTCTGCTTTGTCTCATACCTGGCCCACTTGGTGGGTGATCAGGATCAGTTTGACAGTTTTCTC
AAGGCCTATGTGCATGAATTCAAATTCCGAAGCATCTTAGCCGATGACTTTCTGGACTTCTACTTG
GAATATTTCCCTGAGCTTAAGAAAAAGAGAGTGGATATCATTCCAGGTTTTGAGTTTGATCGATGG
CTGAATACCCCCGGCTGGCCCCCGTACCTCCCTGATCTCTCCCCTGGGGACTCACTCATGAAGCCT
GCTGAAGAGCTAGCCCAACTGTGGGCAGCCGAGGAGCTGGACATGAAGGCCATTGAAGCCGTGGC
CATCTCTCCCTGGAAGACCTACCAGCTGGTCTACTTCCTGGATAAGATCCTCCAGAAATCCCCTCTC
CCTCCTGGGAATGTGAAAAAACTTGGAGACACATACCCAAGTATCTCAAATGCCCGGAATGCAGA
GCTCCGGCTGCGATGGGGCCAAATCGTCCTTAAGAACGACCACCAGGAAGATTTCTGGAAAGTGA
AGGAGTTCCTGCATAACCAGGGGAAGCAGAAGTATACACTTCCGCTGTACCACGCAATGATGGGT
GGCAGTGAGGTGGCCCAGACCCTCGCCAAGGAGACTTTTGCATCCACCGCCTCCCAGCTCCACAGC
AATGTTGTCAACTATGTCCAGCAGATCGTGGCACCCAAGGGCAGTTAG

>SGPR_203_SEQ ID NO:32
ATGGCCGCGCAGTGCTGCTGCCGCCAGGCGCCCGGCGCCGAGGCCGCGCCCGTCCGCCCGCCGCC
CGAGCCGCCGCCCGCCCTGGACGTGGCCTCGGCCTCCAGCGCGCAGCTCTTCCGCCTCCGCCACCT
GCAGCTGGGCCTGGAGCTGCGGCCCGAGGCGCGCGAGTTGGCCGGCTGCCTGGTGCTCGAGCTGT
GCGCGCTGCGGCCCGCGCCCCGCGCGCTCGTGCTCGACGCGCACCCGGCTCTGCGCCTGCACTCAG
CCGCCTTCCGTCGCGCCCCCGCCGCGACGAGAACGCCCTGCGCCTTCGCCTTCTCCGCCCCCGGGC
CGGGGCCCGCGCCGCCGCCCCCGCTGCCCGCCTTCCCCGAGGCGCCCGGCTCCGAGCCCGCCTGCT
GTCCGCTGGCCTTCAGGGTGGACCCGTTCACCGACTACGGCTCCTCGCTCACCGTCACGCTGCCGC
CCGAGCTGCAGGCGCACCAGCCCTTCCAGGTCATCCTGCGGTACACCTCGACCGACGCCCCCGCCA
TCTGGTGGCTGGACCCAGAGCTGACCTATGGCTGCGCCAAGCCCTTCGTCTTCACCCAGGGCCACT
CCGTGTGCAACCGCTCCTTCTTCCCGTGCTTCGACACACCTGCCGTGAAGTGCACCTACTCTGCCGT
CGTCAAGGCGCCATCGGGGGTGCAGGTGCTGATGAGTGCCACCCGGAGTGCATACATGGAGGAAG
AAGGCGTCTTCCACTTCCACATGGAGCACCCCGTGCCCGCCTACCTCGTGGCCCTGGTGGCCGGAG
ACCTCAAGCCGGCAGACATCGGGCCCAGGAGCCGCGTGTGGGCCGAGCCATGCCTCCTGCCCACG
GCCACCAGCAAGCTGTCGGGCGCAGTGGAGCAGTGGCTGAGTGCAGCTGAGCGGCTGTATGGGCC
CTACATGTGGGGCAGGTACGACATTGTCTTCCTGCCACCCTCCTTCCCCATCGTGGCCATGGAGAA
CCCCTGCCTCACCTTCATCATCTCCTCCATCCTGGAGAGCGATGAGTTCCTGGTCATCGATGTCATC
CACGAGGTGGCCCACAGTTGGTTCGGCAACGCTGTCACCAACGCCACGTGGGAAGAGATGTGGCT
GAGCGAGGGCCTGGCCACCTATGCCCAGCGCCGTATCACCACCGAGACCTACGGTGCTGCCTTCAC
CTGCCTGGAGACTGCCTTCCGCCTGGACGCCCTGCACCGGCAGATGAAGCTTCTGGGAGAGGACA
GCCCGGTCAGCAAACTGCAGGTCAAGCTGGAGCCAGGAGTGAATCCCAGCCACCTGATGAACCTG
TTCACCTACGAGAAGGGCTACTGCTTCGTGTACTACCTGTCCCAGCTCTGCGGAGACCCACAGCGC
TTTGATGACTTTCTCCGAGCCTATGTGGAGAAGTACAAGTTCACCAGCGTGGTGGCCCAGGACCTG
CTGGACTCCTTCCTGAGCTTCTTCCCGGAGCTGAAGGAGCAGAGCGTGGACTGCCGGGCAGGGCT
GGAATTCGAGCGCTGGCTCAATGCCACAGGCCCGCCGCTGGCTGAGCCGGACCTGTCTCAGGGAT
CCAGCCTGACCCGGCCCGTGGAGGCCCTTTTCCAGCTGTGGACCGCAGAACCTCTGGACCAGGCA
GCTGCCTCGGCCAGCGCCATTGACATCTCCAAGTGGAGGACCTTCCAGACAGCACTCTTCCTGGAC
CGGCTCCTGGATGGGTCCCCGCTGCCGCAGGAGGTGGTGATGAGCCTGTCCAAGTGCTACTCCTCC
CTGCTGGACTCGATGAACGCTGAGATCCGCATCCGCTGGCTGCAGATTGTGGTCCGCAACGACTAC
TATCCTGACCTCCACAGGGTGCGGCGCTTCCTGGAGAGCCAGATGTCACGCATGTACACCATCCCG
CTGTACGAGGACCTCTGCACCGGTGCCCTCAAGTCCTTCGCGCTGGAGGTCTTCTACCAGACGCAG
GGCCGGCTGCACCCCAACCTGCGCAGAGCCATCCAGCAGATCCTGTCCCAGGGCCTGGGCTCCAG
CACAGAGCCCGCCTCAGAGCCCAGCACGGAGCTGGGCAAGGCTGAAGCAGACACAGACTCGGAC
GCACAGGCCCTGCTGCTTGGGGACGAGGCCCCCAGCAGTGCCATCTCTCTCAGGGACGTCAATGTG
TCTGCCTAG

>SGPR_157_SEQ ID NO:33
ATGGATCCCAAACTCGGGAGAATGGCTGCGTCCCTGCTGGCTGTGCTGCTGCTGCTGCTGGAGCGC
GGCATGTTCTCCTCACCCTCCCCGCCCCCGGCGCTGTTAGAGAAAGTCTTCCAGTACATTGACCTCC

Figure 1V

ATCAGGATGAATTTGTGCAGACGCTGAAGGAGTGGGTGGCCATCGAGAGCGACTCTGTCCAGCCT
GTGCCTCGCTTCAGACAAGAGCTCTTCAGAATGATGGCCGTGGCTGCGGACACGCTGCAGCGCCTG
GGGGCCCGTGTGGCCTCGGTGGACATGGGTCCTCAGCAGCTGCCCGATGGTCAGAGTCTTCCAATA
CCTCCCGTCATCCTGGCCGAACTGGGGAGCGATCCCACGAAAGGCACCGTGTGCTTCTACGGCCAC
TTGGACGTGCAGCCTGCTGACCGGGGCGATGGGTGGCTCACGGACCCCTATGTGCTGACGGAGGT
AGACGGGAAACTTTATGGACGAGGAGCGACCGACAACAAAGGCCCTGTCTTGGCTTGGATCAATG
CTGTGAGCGCCTTCAGAGCCCTGGAGCAAGATCTTCCTGTGAATATCAAATTCATCATTGAGGGGA
TGGAAGAGGCTGGCTCTGTTGCCCTGGAGGAACTTGTGGAAAAAGAAAAGGACCGATTCTTCTCT
GGTGTGGACTACATTGTAATTTCAGATAACCTGTGGATCAGCCAAAGGAAGCCAGCAATCACTTAT
GGAACCCGGGGGAACAGCTACTTCATGGTGGAGGTGAAATGCAGAGACCAGGATTTTCACTCAGG
AACCTTTGGTGGCATCCTTCATGAACCAATGGCTGATCTGGTTGCTCTTCTCGGTAGCCTGGTAGAC
TCGTCTGGTCATATCCTGGTCCCTGGAATCTATGATGAAGTGGTTCCTCTTACAGAAGAGGAAATA
AATACATACAAAGCCATCCATCTAGACCTAGAAGAATACCGGAATAGCAGCCGGGTTGAGAAATT
TCTGTTCGATACTAAGGAGGAGATTCTAATGCACCTCTGGAGGTACCCATCTCTTTCTATTCATGGG
ATCGAGGGCGCGTTTGATGAGCCTGGAACTAAAACAGTCATACCTGGCCGAGTTATAGGAAAATT
TTCAATCCGTCTAGTCCCTCACATGAATGTGTCTGCGGTGGAAAAACAGGTGACACGACATCTTGA
AGATGTGTTCTCCAAAAGAAATAGTTCCAACAAGATGGTTGTTTCCATGACTCTAGGACTACACCC
GTGGATTGCAAATATTGATGACACCCAGTATCTCGCAGCAAAAAGAGCGATCAGAACAGTGTTTG
GAACAGAACCAGATATGATCCGGGATGGATCCACCATTCCAATTGCCAAAATGTTCCAGGAGATC
GTCCACAAGAGCGTGGTGCTAATTCCGCTGGGAGCTGTTGATGATGGAGAACATTCGCAGAATGA
GAAAATCAACAGGTGGAACTACATAGAGGGAACCAAATTATTTGCTGCCTTTTTCTTAGAGATGGC
CCAGCTCCATTAA

>SGPR_154_SEQ ID NO:34
ATGGCTCAGCGGTGCGTTTGCGTGCTGGCCCTGGTGGCTATGCTGCTCCTAGTTTTCCCTACCGTCT
CCAGATCGATGGGCCCGAGGAGCGGGGAGTATCAAAGGGCGTCGCGAATCCCTTCTCAGTTCAGC
AAAGAGGAACGCGTCGCGATGAAAGAGGCACTGAAAGGTGCCATCCAGATTCCAACAGTGACTTT
TAGCTCTGAGAAGTCCAATACTACAGCCCTGGCTGAGTTCGGAAAATACATTCGTAAAGTCTTTCC
TACAGTGGTCAGCACCAGCTTTATCCAGCATGAAGTCGTGGAAGAGTATAGCCACCTGTTCACTAT
CCAAGGCTCGGACCCCAGCTTGCAGCCCTACCTGCTGATGGCTCACTTTGATGTGGTGCCTGCCCC
TGAAGAAGGCTGGGAGGTGCCCCCATTCTCTGGGTTGGAGCGTGATGGCGTCATCTATGGTCGGG
GCACACTGGACGACAAGAACTCTGTGATGGCATTACTGCAGGCCTTGGAGCTCCTGCTGATCAGG
AAGTACATCCCCCGAAGATCTTTCTTCATTTCTCTGGGCCATGATGAGGAGTCATCAGGGACAGGG
GCTCAGAGGATCTCAGCCCTGCTACAGTCAAGGGGCGTCCAGCTAGCCTTCATTGTGGACGAGGG
GGGCTTCATCTTGGATGATTTCATTCCTAACTTCAAGAAGCCCATCGCCTTGATTGCAGTCTCAGAG
AAGGGTTCCATGAACCTCATGCTGCAAGTAAACATGACTTCAGGCCACTCTTCAGCTCCTCCAAAG
GAGACAAGCATTGGCATCCTTGCAGCTGCTGTCAGCCGATTGGAGCAGACACCAATGCCTATCATA
TTTGGAAGCGGGACAGTGGTGACTGTATTGCAGCAACTGGCAAATGAGGTTTATGGAGAGAAATC
CCTTAACCAATGCAATAATCAGGACCACCACGGCACTCACCATATTCAAAGCAGGGTGGCCCAGG
CCACAGTCAACTTCCGGATTCACCCTGGACAGACAGTCCAAGAGGTCCTAGAACTCACGAAGAAC
ATTGTGGCTGATAACAGAGTCCAGTTCCATGTGTTGAGTGCCTTTGACCCCCTCCCCGTCAGCCCTT
CTGATGACAAGGCCTTGGGCTACCAGCTGCTCCGCCAGACCGTACAGTCCGTCTTCCCGGAAGTCA
ATATTACTGCCCCAGTTACTTCTATTGGCAACACAGACAGCCGATTCTTTACAAACCTCACCACTG
GCATCTACAGGTTCTACCCCATCTACATACAGCCTGAAGACTTCAAACGCATCCATGGAGTCAACG
AGAAAATCTCAGTCCAAGCCTATGAGACCCAAGTGAAATTCATCTTTGAGTTGATTCAGAATGCTG
ACACAGACCAGGAGCCAGTTTCTCACCTGCACAAACTGTGA

>SGPR_088_SEQ ID NO:35
ATGGCGGCCCTCACTACCCTGTTTAAGTACATAGATGAAAATCAGGATCGCTACATTAAGAAACTC
GCAAAATGGGTGGCTATCCAGAGTGTGTCTGCGTGGCCGGAGAAGAGAGGCGAAATCAGGAGGA
TGATGGAAGTTGCTGCTGCAGATGTTAAGCAGTTGGGGGGCTCTGTGGAACTGGTGGATATCGGA
AAACAAAAGCTCCCTGATGGCTCGGAGATCCCGCTCCCTCCTATTCTGCTCGGCAGGCTGGGCTCC
GACCCACAGAAGAAGACCGTGTGCATTTACGGGCACCTGGATGTGCAGCCTGCAGCCCTGGAGGA
CGGCTGGGACAGCGAGCCCTTCACCCTGGTGGAGCGAGACGGCAAGCTGTATGGGAGAGGTTCGA
CTGATGATAAGGGCCCGGTGGCCGGCTGGATAAACGCCCTGGAAGCGTATCAGAAAACAGGCCAG
GAGATTCCTGTCAACGTCCGATTCTGCCTCGAAGGCATGGAGGAGTCAGGCTCTGAGGGCCTAGA
CGAGCTGATTTTTGCCCGGAAAGACACATTCTTTAAGGATGTGGACTATGTCTGCATTTCTGACAA
TTACTGGCTGGGAAAGAAGAAGCCCTGCATCACCTACGGCCTCAGGGGCATTTGCTACTTTTTCAT

Figure 1W

CGAGGTGGAGTGCAGCAACAAAGACCTCCATTCTGGGGTGTACGGGGGCTCGGTGCATGAGGCCA
TGACTGATCTCATTTTGCTGATGGGCTCTTTGGTGGACAAGAGGGGGAACATCCTGATCCCCGGCA
TTAACGAGGCCGTGGCCGCCGTCACGGAAGAGGAGCACAAGCTGTACGACGACATCGACTTTGAC
ATAGAGGAGTTTGCCAAGGATGTGGGGGCGCAGATCCTCCTGCACAGCCACAAGAAAGACATCCT
CATGCACCGATGGCGGTACCCGTCTCTGTCCCTCCATGGCATCGAAGGCGCCTTCTCTGGGTCTGG
GGCCAAGACCGTGATTCCCAGGAAGGTGGTTGGCAAGTTCTCCATCAGGCTCGTGCCGAACATGA
CTCCTGAAGTCGTCGGCGAGCAGGTCACAAGCTACCTAACTAAGAAGTTTGCTGAACTACGCAGC
CCCAATGAGTTCAAGGTGTACATGGGCCACGGTGGGAAGCCCTGGGTCTCCGACTTCAGTCACCCT
CATTACCTGGCTGGGAGAAGAGCCATGAAGACAGTTTTTGGTGTTGAGCCAGACTTGACCAGGGA
AGGCGGCAGTATTCCCGTGACCTTGACCTTTCAGGAGGCCACGGGCAAGAACGTCATGCTGCTGCC
TGTGGGGTCAGCGGATGACGGAGCCCACTCCCAGAATGAAAAGCTCAACAGGTATAACTACATAG
AGGGAACCAAGATGCTGGCCGCGTACCTGTATGAGGTCTCCCAGCTGAAGGACTAG

Figure 2A

>SGPr_140_SEQ ID NO:36
MRGLVVFLAVFALSEVNAITRVPLHKGKSLRRALKERRLLEDFLRNHHYAVSRKHSSSGVVASESLTN
YLDCQYFGKIYIGTLPQKFTLVFDTGSPDIWVPSVYCNSDACQNHQRFDPSKSSTQNMGKSLSIQYGTG
SMRGLLGYDTVTVSNIVDPHQTVGLSTQEPGDVFTYSEFDGILGLAYPSLASEYALRLGFRNDQGSMLT
LRAIDLSYYTGSLHWIPMTARILAVHCGQEGPGEGGLDEAILHTFGSVIIDGVVVACDGGCQAILDTGTS
LLVGPGGNILNIQQAIGRTAGQYNEFDIDCGRLSSIPTAVFEIHGKKYPLPPSAYTSQDQGFCTSGFQGD
YSSQQWILGNVFIWEYYSVFDRTNNRVGLAKAV

>SGPr_197_SEQ ID NO:37
MDRCKHVGRLRLAQDHSILNPQKWCCLECATTESVWACLKCSHVACGRYIEDHALKHFEETGHPLAM
EVRDLYVFCYLCKDYVLNDNPEGDLKLLRSSLLAVRGQKQDTPVRRGRTLRSMASGEDVVLPQRAPQ
GQPQMLTALWYRRQRLLARTLRLWFEKSSRGQAKLEQRRQEEALERKKEEARRRRREPAMAPGVTG
LRNLGNTCYMNSILQVLSHLQKFRECFLNLDPSKTEHLFPKATNGKTQLSGKPTNSSATELSLRNDRAE
ACEREGFCWNGRASISRSLELIQNKEPSSKHISLCRELHTLFRVMWSGKWALVSPFAMLHSVWSLIPAF
RGYDQQDAQEFLCELLHKVQQELESEGTTRRILIPFSQRKLTKQVLKVVNTIFHGQLLSQGRWSGRNHR
EKIGVHVVFDQVLTMEPYCCRDMLSSLDKETFAYDLSAVVMHHGKGFGSGHYTAYCYNTEGGEQTQ
GLAITNREYGLSQRELAPPSKAFPLM

>SGPr_005_SEQ ID NO:38
MGPRLIPFLFLFVYPILCRIILRKGKSIRQRMEEQGVLETFLRDHPKADPIAKYYFNNDAVAYEPFTNYL
DSFYFGEISTGTPPQNFLVSLIRVPPICSLPSIYCQSQVCSNHNRFNPSLSSTFRNDGQTYGLSYGSGSLSV
FLGYDTVTVHNIVVNNQEFGLSENEPSDPFYYSDFDGILGMAYPNMAEGNSPTVMQGMLQQSQLTQP
VFSFYFTCQPTRQYCGELILGGVDPNLYSGQIIWTPVSPELYWQIAIEEFAIGNQATGLCSEGCQAIVDTE
TFLLAVPQQYMASFLQATGPQQAQNGDFVVNCSYIQSMPTITFIIGGAQFPLPPSEYVFNNNGYCRLGT
EATCLPSRSGQPLWILGDVFLKEYCSVYDMANNRVGFAFSA

>SGPr_078_SEQ ID NO:39
MQPSSLLPLALCLLAAPASALVRIPLHKFTSIRRTMSEVGGSVEDLIAKGPVSKYSQAVPAVTEGPIPEVL
KNYMDAQYYGEIGIGTPPQCFTVVFDTGSSNLWVPSIHCKLLDIACWIHHKYNSDKSSTYVKNGTSFDI
HYGSGSLSGYLSQDTVSVPCQSASSASALGGVKVERQVFGEATKQPGITFIAAKFDGILGMAYPRISVN
NVLPVFDNLMQQKLVDQNIFSFYLSRDPDAQPGGELMLGGTDSKYYKGSLSYLNVTRKAYWQVHLD
QVEVASGLTLCKEGCEAIVDTGTSLMVGPVDEVRELQKAIGAVPLIQGEYMIPCEKVSTLPAITLKLGG
KGYKLSPEDYTLKVSQAGKTLCLSGFMGMDIPPPSGPLWILGDVFIGRYYTVFDRDNNRVGFAEAARL

>SGPr_084_SEQ ID NO:40
MALLTNLLPLCCLALLALPAQSCGPGRGPVGRRRYARKQLVPLLYKQFVPGVPERTLGASGPAEGRVA
RGSERFRDLVPNYNPDIIFKDEENSGADRLMTERCKERVNALAIAVMNMWPGVRLRVTEGWDEDGHH
AQDSLHYEGRALDITTSDRDRNKYGLLARLAVEAGFDWVYYESRNHVHVSVKADNSLAVRAGGCFP
GNATVRLWSGERKGLRELHRGDWVLAADASGRVVPTPVLLFLDRDLQRRASFVAVETEWPPRKLLLT
PWHLVFAARGPAPAPGDFAPVFARRLRAGDSVLAPGGDALRPARVARVAREEAVGVFAPLTAHGTLL
VNDVLASCYAVLESHQWAHRAFAPLRLLHALGALLPGGAVQPTGMHWYSRLLYRLAEELLG

>SGPr_009_SEQ ID NO:41
MAEKPSNGVLVHMVKLLIKTFLDGIFDDLMENNVLNTDEIHLIGKCLKFVVSNAENLVDDITETAQTA
GKIFREHLWNSKKQLSSIFFSLSAFLEIQGAQPSGKLKLCPHAHFHELKTKRADEIYPVMEKERRTCLGL
NIRNKEFNYLHNRNGSELDLLGMRDLLENLGYSVVIKENLTAQEMETALRQFAAHPEHQSSDSTFLVF
MSHSILNGICGTKHWDQEPDVLHDDTIFEIFNNRNCQSLKDKPKVIIMQACRGNGAGIVWFTTDSGKAG
ADTHGRLLQGNICNDAVTKAHVEKDFIAFKSSTPHNVSWRHETNGSVFISQIIYYFREYSWSHHLEEIFQ
KVQHSFETPNILTQLPTIERLSMTRYFYLFPGN

>SGPr_286_SEQ ID NO:42
QYDLSKARAALLLAVIQGRPGAQHDVEALGGLCWALGFETTVRTDPTAQAFQEELAQFREQLDTCRG
PVSCALVALMAHGGPRGQLLGADGQEVQPEALMQELSRCQVLQGRPKIFLLQACRGGNRDAGVGPTA
LPWYWSWLRAPPSVPSHADVLQIYAEAQGYVAYRDDKGSDFIQTLVEVLRANPGRDLLELLTEVNRR
VCEQEVLGPDCDELRKACLEIRSSLRRRLCLQA

Figure 2B

>SGPr_008_SEQ ID NO:43
MAYYQEPSVETSIIKFKDQDFTTLRDHCLSMGRTFKDETFPAADSSIGQKLLQEKRLSNVIWKRPQDLP
GGPPHFILDDISRFDIQQGGAADCWFLAALGSLTQNPQYRQKILMVQSFSHQYAGIFRFRFWQCGQWV
EVVIDDRLPVQGDKCLFVRPRHQNQEFWPCLLEKAYAKLLGSYSDLHYGFLEDALVDLTGGVITNIHL
HSSPVDLVKAVKTATKAGSLITCATPSGPTDTAQAMENGLVSLHAYTVTGAEQIQYRRGWEEIISLWN
PWGWGEAEWRGRWSDGSQEWEETCDPRKSQLHKKREDGEFWMSCQDFQQKFIAMFICSEIPITLDHG
NTLHEGWSQIMFRKQVILGNTAGGPRNDAQFNFSVQEPMEGTNVVVCVTVAVTPSNLKAEDAKFPLD
FQVILAGSQRFREKFPPVFFSSFRNTVQSSNNKFRRNFTMTYHLSPGNYVVVAQTRRKSAEFLLRIFLKM
PDSDRHLSSHFNLRMKGSPSEHGSQQSIFNRYAQQRLDIDATQLQGLLNQELLTGPPGDMFSLDECRSL
VALMELKVNGRLDQEEFARLWKRLVHYQHVFQKVQTSPGVLLSSDLWKAIENTDFLRGIFISRELLHL
VTLRYSDSVGRVSFPSLVCFLMRLEAMAKTFRNLSKDGKGLYLTEMEWMSLVMYN

>SGPr_198_SEQ ID NO:44
MAAQAAGVSRQRAATQGLGSNQNALKYLGQDFKTLRQQCLDSGVLFKDPEFPACPSALGYKDLGPGS
PQTQGIIWKRPTELCPSPQFIVGGATRTDICQGGLGDCWLLAAIASLTLNEELLYRVVPRDQDFQENYA
GIFHFQFWQYGEWVEVVIDDRLPTKNGQLLFLHSEQGNEFWSALLEKAYAKLNGCYEALAGGSTVEG
FEDFTGGISEFYDLKKPPANLYQIIRKALCAGSLLGCSIDVYSAAEAEAITSQKLVKSHAYSVTGVEEVN
FQGHPEKLIRLRNPWGEVEWSGAWSDDAPEWNHIDPRRKEELDKKVEDGEFWMSLSDFVRQFSRLEIC
NLSPDSLSSEEVHKWNLVLFNGHWTRGSTAGGCQNYPATYWTNPQFKIRLDEVDEDQEESIGEPCCTV
LLGLMQKNRRWRKRIGQGMLSIGYAVYQVPKELESHTDAHLGRDFFLAYQPSARTSTYVNLREVSGR
ARLPPGEYLVVPSTFEPFKDGEFCLRVFSEKKAQALEIGDVVAGNPYEPHPSEVDQEDDQFRRLFEKLA
GKDSEITANALKILLNEAFSKRTDIKFDGFNINTCREMISLLDSNGTGTLGAVEFKTLWLKIQKYLEIYW
ETDYNHSGTIDAHEMRTALRKAGFTLNSQVQQTIALRYACSKLGINFDSFVACMIRLETLFKLFSLLDE
DKDGMVQLSLAEWLCCVLV

>SGPr_210_SEQ ID NO:45
MASSSGRVTIQLVDEEAGVGAGRLQLFRGQSYEAIRAACLDSGILFRDPYFPAGPDALGYDQLGPDSEK
AKGVKWMRPHEFCAEPKFICEDMSRTDVCQGSLGNCWFLAAAASLTLYPRLLRRVVPPGQDFQHGYA
GVFHFQLWQFGRWMDVVVDDRLPVREGKLMFVRSEQRNEFWAPLLEKAYAKLHGSYEVMRGGHM
NEAFVDFTGGVGEVLYLRQNSMGLFSALRHALAKESLVGATAQSDRGEYRTEEGLVKGHAYSITGTH
KVFLGFTKVRLLRLRNPWGCVEWTGAWSDSCPRWDTLPTECRDALLVKKEDGEFWMELRDFLLHFD
TVQICSLSPEVLGPSPEGGGWHVHTFQGRWVRGFNSGGSQPNAETFWTNPQFRLTLLEPDEEDDEDEE
GPWGGWGAAGARGPARGGRTPKCTVLLSLIQRNRRRLRAKGLTYLTVGFHVFQAEGSTGTDNERTHG
FTGHRGAQLAGHTHGPQEASKRYTQNSAEVAPDREADDDGGQGFGDGPWEIDDVISADLQSLQGPYL
PLELGLEQLFQELAGEEEELNASQLQALLSIALEPARAHTSTPREIGLRTCEQLLQCFGHGQSLALHHFQ
QLWGYLLEWQAIFNKFDEDTSGTMNSYELRLALNAAGFHLNNQLTQTLTSRYRDSRLRVDFERFVSC
VAHLTCIFCHCSQHLDGGEGVICLTHRQWMEVATFS

>SGPr_290_SEQ ID NO:46
MSLWPPFRCRWKLAPRYSRRASPQQPQQDFEALLAECLRNGCLFEDTSFPATLSSIGSGSLLQKLPPRLQ
WKRPPELHSNPQFYFAKAKRLDLCQGIVGDCWFLAALQALALHQDILSRVVPLNQSFTEKYAGIFRFW
FWHYGNWVPVVIDDRLPVNEAGQLVFVSSTYKNLFWGALLEKAYAKLSGSYEDLQSGQVSEALVDFT
GGVTMTINLAEAHGNLWDILIEATYNRTLIGCQTHSGEKILENGLVEGHAYTLTGIRKVTCKHRPEYLV
KLRNPWGKVEWKGDWSDSSSKWELLSPKEKILLLRKDNDGEFWMTLQDFKTHFVLLVICKLTPGLLS
QEAAQKWTYTMREGRWEKRSTAGGQRQLLQDTFWKNPQFLLSVWRPEEGRRSLRPCSVLVSLLQKP
RHRCRKRKPLLAIGFYLYRMNKYHDDQRRLPPEFFQRNTPLSQPDRFLKEKEVSQELCLEPGTYLIVPA
YWRPTRSQSSSSGSSPGSTSFMKLAAILVSSSQRRXKTKMKGRMNSSPNSFXKHPEINAVQLQNLLXQ
MTWSSLGSRQPFFSLEACQGILALLDVSFQLNASGTMSIQEFRDLWKQLKLSQKVFHKQDRGSGYLNW
EQLHAAMREAGIMLSDDVCQLMLIRYGGPRLQMDFVSFIHLMLRVENMEGKLAGSWGGPGLPLLPHD
FPPVPSLSTREDSRHPRNSRPGKLWGPPAKCL

Figure 2C

>SGPr_116_SEQ ID NO:47
MVAHINNSRLKAKGVGQHDNAQNFGNQSFEELRAACLRKGELFEDPLFPAEPSSLGFKDLGPNSKNVQ
NISWQRPKDIINNPLFIMDGISPTDICQGILGDCWLLAAIGSLTTCPKLLYRVVPRGQSFKKNYAGIFHFQI
WQFGQWVNVVVDDRLPTKNDKLVFVHSTERSEFWSALLEKAYAKLSGSYEALSGGSTMEGLEDFTG
GVAQSFQLQRPPQNLLRLLRKAVERSSLMGCSIEVTSDSELESMTDKMLVRGHAYSVTGLQDVHYRG
KMETLIRVRNPWGRIEWNGAWSDSAREWEEVASDIQMQLLHKTEDGEFWMSYQDFLNNFTLLEICNL
TPDTLSGDYKSYWHTTFYEGSWRRGSSAGGCRNHPGTFWTNPQFKISLPEGDDPEDDAEGNVVVCTC
LVALMQKNWRHARQQGAQLQTIGFVLYAVPKEFQNIQDVHLKKEFFTKYQDHGFSEIFTNSREVSSQL
RLPPGEYIIIPSTFEPHRDADFLLRVFTEKHSESWELDEVNYAEQLQEEKVSEDDMDQDFLHLFKIVAGE
GKEIGVYELQRLLNRMAIKFKSFKTKGFGLDACRCMINLMDKDGSGKLGLLEFKILWKKLKKWMDIF
RECDQDHSGTLNSYEMRLVIEKAGIKLNNKVMQVLVARYADDDLIIDFDSFISCFLRLKTMFTFFLTMD
PKNTGHICLSLEQWLQMTMWG

>SGPr_003_SEQ ID NO:48
MRAGRGATPARELFRDAAFPAADSSLFCDLSTPLAQFREDITWRRPQEICATPRLFPDDPREGQVKQGL
LGDCWFLCACAALQKSRHLLDQVIPPGQPSWADQEYRGSFTCRIWQFGRWVEVTTDDRLPCLAGRLC
FSRCQREDVFWLPLLEKVYAKVHGSYEHLWAGQVADALVDLTGGLAERWNLKGVAGSGGQQDRPG
RWEHRTCRQLLHLKDQCLISCCVLSPRAGARELGEFHAFIVSDLRELQGQAGQCILLLRIQNPWGRRC
WQGLWREGGEGWSQVDAAVASELLSQLQEGEFWVEEEEFLREFDELTVGYPVTEAGHLQSLYTERLL
CHTRALPGAWVKGQSAGGCRNNSGFPSNPKFWLRVSEPSEVYIAVLQRSRLHAADWAGRARALVGDS
HTSWSPASIPGKHYQAVGLHLWKVEKRRVNLPRVLSMPPVAGTACHAYDREVHLRCELSPGYYLAVP
STFLKDAPGEFLLRVFSTGRVSLRSQRVEGARTHPHCCCRSRC

>SGPr_016_SEQ ID NO:49
MFLLLVLLTGLGGMHADLNPHKIFLQTTIPEKISSSDAKTDPEHNVILIIFLLEIMFLLFLPRSILSSASVINS
YDENDIRHSKPLLVQMDCIYNGYVAGIPNSLVTLSVCSGLRLGTMQLKNISYGIEPMEAKTDFIKLFPRY
IEMHIVVDKNLVKTIKSIWXMFSQLKTSITLSSLELWSDENKISTNGVADDVLQRFLSWKQKFMSQKSN
IVAYLLMXYSGGVKDFNICSLDDFKYISSHNGLTCLQTNPLEMPTYTHRRICGNGLLEGSEECDCGTKD

>SGPr_352_SEQ ID NO:50
MAPACQILRWALALGLGLMFEVTHAFRSQDEFLSSLESYEIAFPTRVDHNGALLAFSPPPPRRQRRGTG
ATAESRLFYKVASPSTHFLLNLTRSSRLLAGHVSVEYWTREGLAWQRAARPHCLYAGHLQGQASSSH
VAISTCGGLHGLIVADEEEYLIEPLHGGPKGSRSPEESGPHVVYKRSSLRHPHLDTACGVRDEKPWKGR
PWWLRTLKPPPARPLGNETERGQPGLKRSVSRERYVETLVVADKMMVAYHGRRDVEQYVLAVMNIV
AKLFQDSSLGSTVNILVTRLILLTEDQPTLEITHHAGKSLDSFCKWQKSIVNHSGHGNAIPENGVANHDT
AVLITRYDICIYKNKPCGTLGLAPVGGMCERERSCSVNEDIGLATAFTIAHEIGHTFGMNHDGVGNSCG
ARGQDPAKLMAAHITMKTNPFVWSSCSRDYITSFLDSGLGLCLNNRPPRQDFVYPTVAPGQAYDADEQ
CRFQHGVKSRQCKYGEVCSELWCLSKSNRCITNSIPAAEGTLCQTHTIDKGWCYKRVCVPFGSRPEGV
DGAWGPWTPWGDCSRTCGGGVSSSSRHCDSPRPTIGGKYCLGERRRHRSCNTDDCPPGSQDFREVQCS
EFDSIPFRGKFYKWKTYRGGGVKACSLTCLAEGFNFYTERAAAVVDGTPCRPDTVDICVSGECKHVGC
DRVLGSDLREDKCRVCGGDGSACETIEGVFSPASPGAGYEDVVWIPKGSVHIFIQDLNLSLSHLALKGD
QESLLLEGLPGTPQPHRLPLAGTTFQLRQGPDQVQSLEALGPINASLIVMVLARTELPALRYRFNAPIAR
DSLPPYSWHYAPWTKCSAQCAGGSQVQAVECRNQLDSSAVAPHYCSAHSKLPKRQRACNTEPCPPDW
VVGNWSLCSRSCDAGVRSRSVVCQRRVSAAEEKALDDSACPQPRPPVLEACHGPTCPPEWAALDWSE
CTPSCGPGLRHRVVLCKSADHRATLPPAHCSPAAKPPATMRCNLRRCPPARWVAGEWGECSAQCGVG
QRQRSVRCTSHTGQASHECTEALRPPTTQQCEAKCDSPTPGDGPEECKDVNKVAYCPLVLKFQFCSRA
YFRQMCCKTCQGH

Figure 2D

>SGPr_050_SEQ ID NO:51
MKPRARGWRGLAALWMLLAQVAEQAPACAMGPAAAAPGSPSVPRPPPPAERPGWMEKGEYDLVSA
YEVDHRGDYVSHEIMHHQRRRRAVAVSEVESLHLRLKGPRHDFHMDLRTSSSLVAPGFIVQTLGKTGT
KSVQTLPPEDFCFYQGSLRSHRNSSVALSTCQGLSGMIRTEEADYFLRPLPSHLSWKLGRAAQGSSPSH
VLYKRSTEPHAPGASEVLVTSRTWELAHQPLHSSDLRLGLPQKQHFCGRRKKYMPQPPKEDLFILPDEY
KSCLRHKRSLLRSHRNEELNVETLVVVDKKMMQNHGHENITTYVLTILNMVSALFKDGTIGGNINIAIV
GLILLEDEQPGLVISHHADHTLSSFCQWQSGLMGKDGTRHDHAILLTGLDICSWKNEPCDTLGFAPISG
MCSKYRSCTINEDTGLGLAFTIAHESGHNFGMIHDGEGNMCKKSEGNIMSPTLAGRNGVFSWSPCSRQ
YLHKFLSTAQAICLADQPKPVKEYKYPEKLPGELYDANTQCKWQFGEKAKLCMLDFKKDICKALWCH
RIGRKCETKFMPAAEGTICGHDMWCRGGQCVKYGDEGPKPTHGHWSDWSSWSPCSRTCGGGVSHRS
RLCTNPKPSHGGKFCEGSTRTLKLCNSQKCPRDSVDFRAAQCAEHNSRRFRGRHYKWKPYTQVEDQD
LCKLYCIAEGFDFFFSLSNKVKDGTPCSEDSRNVCIDGICERVGCDNVLGSDAVEDVCGVCNGNNSACT
IHRGLYTKHHHTNQYYHMVTIPSGARSIRIYEMNVSTSYISVRNALRRYYLNGHWTVDWPGRYKFSGT
TFDYRRSYNEPENLIATGPTNETLIVELLFQGRNPGVAWEYSMPRLGTEKQPPAQPSYTWAIVRSECSV
SCGGGQMTVREGCYRDLKFQVNMSFCNPKTRPVTGLVPCKVSACPPSWSVGNWSACSRTCGGGAQS
RPVQCTRRVHYDSEPVPASLCPQPAPSSRQACNSQSCPPAWSAGPWAECSHTCGKGWRKRAVACKST
NPSARAQLLPDAVCTSEPKPRMHEACLLQRCHKPKKLQWLVSAWSQCSVTCERGTQKRFLKCAEKYV
SGKYRELASKKCSHLPKPSLELERACAPLPCPRHPPFAAAGPSRGSWFASPWSQCTASCGGGVQTRSVQ
CLAGGRPASGCLLHQKPSASLACNTHFCPIAEKKDAFCKDYFHWCYLVPQHGMCSHKFYGKQCCKTC
SKSNL

>SGPr_282_SEQ ID NO:52
MRQAEARVTLRAPLLLLGLWVLLTPVRCSQGHPSWHYASSKVVIPRKETHHGKDLQFLGWLSYSLHF
GGQRHIIHMRRKHLLWPRHLLVTTQDDQGALQMDDPYIPPDCYYLSYLEEVPLSMVTVDMCCGGLRG
IMKLDDLAYEIKPLQDSRRLEHVSQIVAEPNATGPTFRDGDNEETNPLFSEANDSMNPRISNWLYSSHR
GNIKGHVQCSNSYCRVDDNITTCSKEVVQMFSLSDSIVQNIDLRYYIYLLTIYNNCDPAPVNDYRVQSA
MFTYFRTTFFDTFRVHSPTLLIKEAPHECNYEPQRYSFCTHLGLLHIGTLGRHYLLVAVITTQTLMRSTG
EKYDDNYCTCQKRAFCIMQQYPGMTDAFSNCSYGHAQNCFVHSARCVFETLAPVYNETMTMVRCGN
LIADGREECDCGSFKQCYASYCCRSDCRLTPGSICHIGECCTNCSYSPPGTLCRPIQNICDLPEYCHGTTV
TCPANFYMQDGTPCTEEGYCYHGNCTDRNVLCKVIFGVSAEEAPEVCYDINLESYRFGHCTRRQTALN
NQACAGIDKFCGRLQCTSVTHLPRLQEHVSFHHSVTGGFQCFGLDDHRATDTTDVGCVIDGTPCVHGN
FCNNTRCNATITSLGYDCRPEKCSHRGVCNNRRNCHCHIGWDPPLCLRRGAGGSVDSGPPPKITRSVKQ
SQQSVMYLRVVFGRIYTFIIALLFGMATNVRTIRTTTVKGWTVTNPE

>SGPr_046_SEQ ID NO:53
MVEKHGKGNVTTYILTVMNMVSGLFKDGTIGSDINVVVVSLILLEQEPGGLLINHHADQSLNSFCQWQ
SALIGKNGKRHDHAILLTGFDICSWKNEPCDTLGFAPISGMCSKYRSCTINEDTGLGLAFTIAHESGHNF
GMIHDGEGNPCRKAEGNIMSPTLTGNNGVFSWSSCSRQYLKKFLSTPQAGCLVDEPKQAGQYKYPDK
LPGQIYDADTQCKWQFGAKAKLCSLGFVKDICKSLWCHRVGHRCETKFMPAAEGTVCGLSMWCRQG
QCVKFGELGPRPIHGQWSAWSKWSECSRTCGGGVKFQERHCNNPKPQYGGLFCPGSSRIYQLCNINPC
NENSLDFRAQQCAEYNSKPFRGWFYQWKPYTKVEEEDRCKLYCKAENFEFFFAMSGKVKDGTPCSPN
KNDVCIDGVCELVGCDHELGSKAVSDACGVCKGDNSTCKFYKGLYLNQHKANEYYPVVLIPAGARSI
EIQELQVSSSYLAVRSLSQKYYLTGGWSIDWPGEFPFAGTTFEYQRSFNRPERLYAPGPTNETLVFEILM
QGKNPGIAWKYALPKVMNGTPPATKRPAYTWSIVQSECSVSCGGGYINVKAICLRDQNTQVNSSFCSA
KTKPVTEPKICNAFSCPAYWMPGEWSTCSKACAGGQQSRKIQCVQKKPFQKEEAVLHSLCPVSTPTQV
QACNSHACPPQWSLGPWSQCSKTCGRGVRKRELLCKGSAAETLPERKRELLCKGSAAETLPESQCTSL
PRPELQEGCVLGRCPKNSRLQWVASSWSECSATCGLGVRKREMKCSEKGFQGKLITFPERRCRNIKKP
NLDLEETCNRRACPAHPVYNMVAGWYSLPWQQCTVTCGGGVQTRSVHCVQQGRPSSSCLLHQKPPV
LRACNTNFCPAPEKREDPSCVDFFNWCHLVPQHGVCNHKFYGKQCCKSCTRKI

Figure 2E

>SGPr_060_SEQ ID NO:54
MDGRGAFWTVAIPRARQEGLGRLGLPFPVKRTPPAPQNPGGSTQAPQRVVGKSHSGIRMPAKSRNLRL
ESKLNRKVVKYKWGKQGSGAGRELVPAFPTNAGLGRRDRCRPPPAGGDVASHGLPGSGVGYSCNQR
EEGLRGGCGGIPHVPLFLSPLPLDASGQRPSSTYRQSLRRGLGTRAHQSPANEIPELGDLRGSRLAQEPA
VLFGLRPSISKRGLLARRLWAQPMLLSGWVVSTTTTIITVTVTFTPTGLLCVKHSRGPLQPTCQESAPEN
RVGKALITFSKGWRASLRLAPPPSALLLRRHGPGGLPVPGTMCDGALLPPLVLPVLLLLVWGLDPGTGS
APSHSPLHPASCGYLPSAFSRRPGGPGAAAGPLTAPERRRRGPRPEYGNRVAPWQARRRRVSARRCAA
PFREVLARLRRRPSPGGAGQRGAVGDAAADVEVVLPWRVRPDDVHLPPLPAAPGPRRRRRPRTPPAAP
RARPGERALLLHLPAFGRDLYLQLRRDLRFLSRGFEVEEAGAARRRGRPAELCFYSGRVLGHPGSLVSL
SACGAAGGLVGLIQLGQEQVLIQPLNNSQGPFSGREHLIRRKWSLTPSPSAEAQRPEQLCKVLTEKKKP
TWGRPSRDWRERRNAIRLTSEHTVETLVVADADMVQYHGAEAAQRFILTVMNMVYNMFQHQSLGIKI
NIQVTKLVLLRQRPAKLSIGHHGERSLESFCHWQNEEYGGARYLGNNQVPGGKDDPPLVDAAVFVTR
TDLCVHKDEPCDTVGIAYLGGVCSAKRKCVLAEDNGLNLAFTIAHELGHNLGMNHDDDHSSCAGRSH
IMSGEWVKGRNPSDLSWSSCSRDDLENFLKSKVSTCLLVTDPRSQHTVRLPHKLPGMHYSANEQCQIL
FGMNATFCRNMEHLMCAGLWCLVEGDTSCKTKLDPPLDGTECGADKWCRAGECVSKTPIPEHVDGD
WSPWGAWSMCSRTCGTGARFRQRKCDNPPPGPGGTHCPGASVEHAVCENLPCPKGLPSFRDQQCQAH
DRLSPKKKGLLTAVVVDDKPCELYCSPLGKESPLLVADRVLDGTPCGPYETDLCVHGKCQKIGCDGIIG
SAAKEDRCGVCSGDGKTCHLVKGDFSHARGTGYIEAAVIPAGARRIRVVEDKPAHSFLAVVVDDKPCE
LYCSPLGKESPLLVADRVLDGTPCGPYETDLCVHGKCQKIGCDGIIGSAAKEDRCGVCSGDGKTCHLV
KGDFSHARGTGYIEAAVIPAGARRIRVVEDKPAHSFLALKDSGKGSINSDWKIELPGEFQIAGTTVRYV
RRGLWEKISAKGPTKLPLHLMVLLFHDQDYGIHYEYTVPVNRTAENQSEPEKPQDSLFIWTHSGWEGC
SVQCGGGEWPWSMTCWVWGFAEGRRKASVASTQSVRHLQPVAPWEFNHIPPKISLQNTWTESSQLPH

>SGPr_068_SEQ ID NO:55
MAPLRALLSYLLPLHCALCAAAGSRTPELHLSGKLSDYGVTVPCSTDFRGRFLSHVVSGPAAASAGSM
VVDTPPTLPRHSSHLRVARSPLHPGGTLWPGRVGRHSLYFNVTVFGKELHLRLRPNRRLVVPGSSVEW
QEDFRELFRQPLRQECVYTGGVTGMPGAAVAISNCDGLAGLIRTDSTDFFIEPLERGQQEKEASGRTHV
VYRREAVQQEWAEPDGDLHNEAFGLGDLPNLLGLVGDQLGDTERKRRHAKPGSYSIEVLLVVDDSVV
RFHGKEHVQNYVLTLMNIVDEIYHDESLGVHINIALVRLIMVGYRQSLSLIERGNPSRSLEQVCRWAHS
QQRQDPSHAEHHDHVVFLTRQDFGPSGYAPVTGMCHPLRSCALNHEDGFSSAFVIAHETGHVLGMEH
DGQGNGCADETSLGSVMAPLVQAAFHRFHWSRCSKLELSRYLPSYDCLLDDPFDPAWPQPPELPGINY
SMDEQCRFDFGSGYQTCLAFRTFEPCKQLWCSHPDNPYFCKTKKGPPLDGTECAPGKWCFKGHCIWK
SPEQTYGQDGGWSSWTKFGSCSRSCGGGVRSRSRSCNNPSPAYGGRPCLGPMFEYQVCNSEECPGTYE
DFRAQQCAKRNSYYVHQNAKHSWVPYEPDDDAQKCELICQSADTGDVVFMNQVVHDGTRCSYRDP
YSVCARGECVPVGCDKEVGSMKADDKCGVCGGDNSHCRTVKGTLGKASKQAGALKLVQIPAGARHI
QIEALEKSPHRIVVKNQVTGSFILNPKGKEATSRTFTAMGLEWEDAVEDAKESLKTSGPLPEAIAILALP
PTEGGPRSSLAYKYVIHEDLLPLIGSNNVLLEEMDTYEWALKSWAPCSKACGGGIQFTKYGCRRRRDH
HMVQRHLCDHKKRPKPIRRRCNQHPCSQPVWVTEEWGACSRSCGKLGVQTRGIQCLMPLSNGTHKV
MPAKACAGDRPEARRPCLRVPCPAQWRLGAWSQCSATCGEGIQQRQVVCRTNANSLGHCEGDRPDT
VQVCSLPACGAEPCTGDRSVFCQMEVLDRYCSIPGYHRLCCVSCIKKASGPNPGPDPGPTSLPPFSTPGS
PLPGPQDPADAAEPPGKPTGSEDHQHGRATQLPGALDTSSPGTQHPFAPETPIPGASWSISPTTPGGLPW
GWTQTPTPVPEDKGQPGEDLRHPGTSLPAASPVT

Figure 2F

>SGPr_096_SEQ ID NO:56
MQFVSWATLLTLLVRDLAEMGSPDAAAAVRKDRLHPRQVKLLETLSEYEIVSPIRVNALGEPFPTNVH
FKRTRRSINSATDPWPAFASSSSSSTSSQAHYRLSAFGQQFLFNLTANAGFIAPLFTVTLLGTPGVNQTKF
YSEEEAELKHCFYKGYVVNTNSEHTAVISLCSGMLGTFRSHDGDYFIEPLQSMDEQEDEEEQNKPHIIYR
RSAPQREPSTGRHACDTSEHKNRHSKDKKKTRARKWGERINLAGDVAALNSGLATEAFSAYGNKTDN
TREKRTHRRTKRFLSYPRFVEVLVVADNRMVSYHGENLQHYILTLMSIVASIYKDPSIGNLINIVIVNLIV
IHNEQDGPSISFNAQTTLKNFCQWQHSKNSPGGIHHDTAVLLTRQDICRAHDKCDTLGLAELGTICDPY
RSCSISEDSGLSTAFTIAHELGHVFNMPHDDNNKCKEEGVKSPQHVMAPTLNFYTNPWMWSKCSRKYI
TEFLDTGYGECLLNEPESRPYPLPVQLPGILYNVNKQCELIFGPGSQVCPYMMQCRRLWCNNVNGVHK
GCRTQHTPWADGTECEPGKHCKYGFCVPKEMDVPVTDGSWGSWSPFGTCSRTCGGGIKTAIRECNRP
EPKNGGKYCVGRRMKFKSCNTEPCLKQKRDFRDEQCAHFDGKHFNINGLLPNVRWVPKYSGILMKDR
CKLFCRVAGNTAYYQLRDRVIDGTPCGQDTNDICVQGLCRQAGCDHVLNSKARRDKCGVCGGDNSS
CKTVAGTFNTVHYGYNTVVRIPAGATNIDVRQHSFSGETDDDNYLALSSSKGEFLLNGNFVVTMAKRE
IRIGNAVVEYSGSETAVERINSTDRIEQELLLQVLSVGKLYNPDVRYSFNIPIEDKPQQFYWNSHGPWQA
CSKPCQGERKRKLVCTRESDQLTVSDQRCDRLPQPGHITEPCGTDCDLRWHVASRSECSAQCGLGYRT
LDIYCAKYSRLDGKTEKVDDGFCSSHPKPSNREKCSGECNTGGWRYSAWTECSKSCDGGTQRRRAICV
NTRNDVLDDSKCTHQEKVTIQRCSEFPCPQWKSGDWSECLVTCGKGHKHRQVWCQFGEDRLNDRMC
DPETKPTSMQTCQQPECASWQAGPWGQCSVTCGQGYQLRAVKCIIGTYMSVVDDNDCNAATRPTDT
QDCELPSCHPPPAAPETRRSTYSAPRTQWRFGSWTPCSATCGKGTRMRYVSCRDENGSVADESACATL
PRPVAKEECSVTPCGQWKALDWSSCSVTCGQGRATRQVMCVNYSDHVIDRSECDQDYIPETDQDCSM
SPCPQRTPDSGLAQHPFQNEDYRPRSASPSRTHVLGGNQWRTGPWGACSSTCAGGSQRRVVVCQDEN
GYTANDCVERIKPDEQRACESGPCPQWAYGNWGECTKLCGGGIRTRLVVCQRSNGERFPDLSCEILDK
PPDREQCNTHACPHDAAWSTGPWSSCSVSCGRGHKQRNVYCMAKDGSHLESDYCKHLAKPHGHRKC
RGGRCPKWKAGAWSQCSVSCGRGVQQRHVGCQIGTHKIARETECNPYTRPESERDCQGPRCPLYTWR
AEEWQECTKTCGEGSRYRKVVCVDDNKNEVHGARCDVSKRPVDRESCSLQPCEYVWITGEWSECSVT
CGKGYKQRLVSCSEIYTGKENYEYSYQTTINCPGTQPPSVHPCYLRDCPVSATWRVGNWGSCSVSCGV
GVMQRSVQCLTNEDQPSHLCHTDLKPEERKTCRNVYNCELPQNCKEVKRLKGASEDGEYFLMIRGKL
LKIFCAGMHSDHPKEYVTLVHGDSENFSEVYGHRLHNPTECPYNGSRRDDCQCRKDYTAAGFSSFQKI
RIDLTSMQIITTDLQFARTSEGHPVPFATAGDCYSAAKCPQGRFSINLYGTGLSLTESARWISQGNYAVS
DIKKSPDGTRVVGKCGGYCGKCTPSSGTGLEVRVL

>SGPr_119_SEQ ID NO:57
MWVAKWLTGLLYHLSLFITRSWEVDFHPRQEALVRTLTSYEVVIPERVNEFGEVFPQSHHFSRQKRSSE
ALEPMPFRTHYRFTAYGQLFQLNLTADASFLAAGYTEVHLGTPERGAWESDAGPSDLRHCFYRGQVN
SQEDYKAVVSLCGGLTGTFKGQNGEYFLEPIMKADGNEYEDGHNKPHLIYRQDLNNSFLQTLKYCSVS
ESQIKETSLPFHTYSNMNEDLNVMKERVLGHTSKNVPLKDERRHSRKKRLISYPRYIEIMVTADAKVVS
AHGSNLQNYILTLMSIVATIYKDPSIGNLIHIVVVKLVMIHREEEGPVINFDGATTLKNFCSWQQTQNDL
DDVHPSHHDTAVLITREDICSSKEKCNMLGLSYLGTICDPLQSCFINEEKGLISAFTIAHELGHTLGVQH
DDNPRCKEMKVTKYHVMAPALSFHMSPWSWSNCSRKYVTEFLDTGYGECLLDKPDEEIYNLPSELPG
SRYDGNKQCELAFGPGSQMCPHIENICMHLWCTSTEKLHKGCFTQHVPPADGTDCGPGMHCRHGLCV
NKETETRPVNGEWGPWEPYSSCSRTCGGGIESATRRCNRPEPRNGGNYCVGRRMKFRSCNTDSCPKGT
QDFREKQCSDFNGKHLDISGIPSNVRWLPRYSGIGTKDRCKLYCQVAGTNYFYLLKDMVEDGTPCGTE
THDICVQGQCMAAGCDHVLNSSAKIDKCGVCGGDNSSCKTITGVFNSSHYGYNVVVKIPAGATNVDIR
QYSYSGQPDDSYLALSDAEGNFLFNGNFLLSTSKKEINVQGTRTVIEYSGSNNAVERINSTNRQEKELIL
QVLCVGNLYNPDVHYSFNIPLEERSDMFTWDPYGPWEGCTKMCQGLQRRNITCIHKSDHSVVSDKEC
DHLPLPSFVTQSCNTDCELRWHVIGKSECSSQCGQGYRTLDIHCMKYSIHEGQTVQVDDHYCGDQLKP
PTQELCHGNCVFTRWHYSEWSQCSRSCGGGERSRESYCMNNFGHRLADNECQELSRVTRENCNEFSC
PSWAASEWSECLVTCGKGTKQRQVWCQLNVDHLSDGFCNSSTKPESLSPCELHTCASWQVGPWGPCT
TTCGHGYQMRDVKCVNELASAVLEDTECHEASRPSDRQSCVLTPCSFISKLETALLPTVLIKKMAQWR
HGSWTPCSVSCGRGTQARYVSCRDALDRIADESYCAHLPRPAEIWDCFTPCGEWQAGDWSPCSASCG
HGKTTRQVLCMNYHQPIDENYCDPEVRPLMEQECSLAACPPAHSHFPSSPVQPSYYLSTNLPLTQKLED
NENQVVHPSVRGNQWRTGPWGSCSSSCSGGLQHRAVVCQDENGQSASYCDAASKPPELQQCGPGPCP
QWNYGNWGECSQTCGGGIKSRLVICQFPNGQILEDHNCEIVNKPPSVIQCHMHACPADVSWHQEPWTS
EDLKVKLLPQRTIILWELMKNIFCHGKHSHMYLINVVTDHLLYPRHCDPETIETYFLSLWSLQFTWGDL
KYYKNSL

Figure 2G

>SGPr_143_SEQ ID NO:58
MGRPVPASAPPRPQLLRTLDIQVALTGLEVRRRRPEAAPGTGRPQSSLGGAGVASRCLEAEELTAMGW
RPRRARGTPLLLLLLLLLLWPVPGAGVLQGHIPGQPVTPHWVLDGQPWRTVSLEEPVSKPDMGLVALE
AEGQELLLELEKNHRLLAPGYIETHYGPDGQPVVLAPNHTDHCHYQGRVRGFPDSWVVLCTCSGMSG
LITLSRNASYYLRPWPPRGSKDFSTHEIFRMEQLLTWKGTCGHRDPGNKAGMTSLPGGPQSRVRREAR
RTRKYLELYIVADHTLFLTRHRNLNHTKQRLLEVANYVDQLLRTLDIQVALTGLEVWTERDRSRVTQD
ANATLWAFLQWRRGLWAQRPHDSAQLLTGRAFQGATVGLAPVEGMCRAESSGGVSTDHSELPIGAA
ATMAHEIGHSLGLSHDPDGCCVEAAAESGGCVMAAATGVVYEHPFPRVFSACSRRQLRAFFRKGGGA
CLSNAPDPGLPVPPALCGNGFVEAGEECDCGPGQECRDLCCFAHNCSLRPGAQCAHGDCCVRCLLKPA
GALCRQAMGDCDLPEFCTGTSSHCPPDVYLLDGSPCARGSGYCWDGACPTLEQQCQQLWGPGSHPAP
EACFQVVNSAGDAHGNCGQDSEGHFLPCAGRDALCGKLQCQGGKPSLLAPHMVPVDSTVHLDGQEV
TCRGALALPSAQLDLLGLGLVEPGTQCGPRMVCQSRRCRKNAFQELQRCLTACHSHGVCNSNHNCHC
APGWAPPFCDKPGFGGSMDSGPVQAENHDTFLLAMLLSILLPLLPGAGLAWCCYRLPGAHLQRCSWG
CRRDPACSGPKDGPHRDHPLGGVHPMELGPTATGQPWPLDPENSHEPSSHPEKPLPAVSPDPQADQVQ
MPRSCLW

>SGPr_164_SEQ ID NO:59
HGDRGSGRRARPSPFPQRGGALPGAMLLLGILTLAFAGRTAGGSEPEREVVVPIRLDPDINGRRYYWRG
PEDSGDQGLIFQITAFQEDFYLHLTPDAQFLAPAFSTEHLGVPLQGLTGGSSDLRRCFYSGDVNAEPDSF
AAVSLCGGLRGAFGYRGAEYVISPLPNASAPAAQRNSQGAHLLQRRGVPGGPSGDPTSRCGVASGWN
PAILRALDPYKPRRAGFGESRSRRRSGRAKRFVSIPRYVETLVVADESMVKFHGADLEHYLLTLLATAA
RLYRHPSILNPINIVVVKVLLLRDRDSGPKVTGNAALTLRNFCAWQKKLNKVSDKHPEYWDTAILFTR
QDLCGATTCDTLGMADVGTMCDPKRSCSVIEDDGLPSAFTTAHELGHVFNMPHDNVKVCEEVFGKLR
ANHMMSPTLIQIDRANPWSACSAAIITDFLDSGHGDCLLDQPSKPISLPEDLPGASYTLSQQCELAFGVG
SKPCPYMQYCTKLWCTGKAKGQMVCQTRHFPWADGTSCGEGKLCLKGACVERHNLNKHRVDGSWA
KWDPYGPCSRTCGGGVQLARRQCTNPTPANGGKYCEGVRVKYRSCNLEPCPSSASGKSFREEQCEAF
NGYNHSTNRLTLAVAWVPKYSGVSPRDKCKLICRANGTGYFYVLAPKVVVDGTLCSPDSTSVCVQGK
CIKAGCDGNLGSKKRFDKCGVCGGDNKSCKKVTGLFTKPMHGYNFVVAIPAGASSIDIRQRGYKGLIG
DDNYLALKNSQGKYLLNGHFVVSAVERDLVVKGSLLRYSGTGTAVESLQASRPILEPLTVEVLSVGKM
TPPRVRYSFYLPKEPREDKSSHPPHPRGGGPSVLHNSVLSLSNQVEQPDDRPPARWVAGSWGPCSASCG
SGLQKRAVDCRGSAGQRTVPACDAAHRPVETQACGEPCPTWELSAWSPCSKSCGRGFQRRSLKCVGH
GGRLLARDQCNLHRKPQELDFCVLRPC

>SGPr_281_SEQ ID NO:60
APDSHLLLLPPLPAGVPVEWDRFRAAVRPRPRGVGSRVSCALAPGAGGPGWRQRGQRGPGLGARRW
GRRKRPGAGCRQLTRGALLWLRCLWRSPWRADQSPGSGPRRRRRVRRTRSFESQELPRGSSGAAALSP
GAPASWQPPPPPQPPPSPPPAQHAEPDGDEVLLRIPAFSRDLYLLLRRDGRFLAPRFAVEQRPNPGPGPT
GAASAPQPPAPPDAGCFYTGAVLRHPGSLASFSTCGGGLVFNLFQHKSLGVQVNLRVIKLILLHETPPEL
YIGHHGEKMLESFCKWQHEEFGKKNDIHLEMSTNWGEDMTSVDAAILITRKDFCVHKDEPCDTVGIA
YLSGMCSEKRKCIIAEDNGLNLAFTIAHEMGHNMGINHDNDHPSCADGLHIMSGEWIKGQNLGDVSW
SRCSKEDLERFLRSKASNCLLQTNPQSVNSVMVPSKLPGMTYTADEQCQILFGPLASFCQEMQHVICTG
LWCKVEGEKECRTKLDPPMDGTDCDLGKWCKAGECTSRTSAPEHLAGEWSLWSPCSRTCSAGISSRE
RKCPGLDSEARDCNGPRKQYRICENPPCPAGLPGFRDWQCQAYSVRTSPPKHILQWQAVLDEEKPCAL
FCSPVGKEQPILLSEKVMDGTSCGYQGLDICANGRCQKVGCDGLLGSLAREDHCGVCNGNGKSCKIIK
GDFNHTRGAGYVEVLVIPAGARRIKVVEEKPAHSYLALRDAGKQSINSDWKIEHSGAFNLAGTTVHYV
RRGLWEKISAKGPTTAPLHLLVLLFQDQNYGLHYEYTIPSDPLPENQSSKAPEPLFMWTHTSWEDCDA
TCGGGERKTTVSCTKIMSKNISIVDNEKCKYLTKPEPQIRKCNEQPCQTRWMMTEWTPCSRTCGKGMQ
SRQVACTQQLSNGTLIRARERDCIGPKPASAQRCEGQDCMTVWEAGVWSECSVKCGKGIRHRTVRCT
NPRKKCVLSTRPREAEDCEDYSKCYVWRMGDWSKCSITCGKGMQSRVIQCMHKITGRHGNECFSSEK
PAAYRPCHLQPCNEKINVNTITSPRLAALTFKCLGDQWPVYCRVIREKNLCQDMRWYQRCCETCRDFY
AQKLQQKS

>SGPr_075_SEQ ID NO:61
YDYWGSDSMIVTNKVIEIVGLANSMFTQFKVTIVLSSLELWSDENKISTVGEADELLQKFLEWKQSYLN
LRPHDIAYLLXYPKEITLEAFAVIVTQMLALSLGISYDDPKKCQCSESTCIMNPEV

Figure 2H

>SGPr_292_SEQ ID NO:62
MLAASIFRPTLLLCWLAAPWPTQPESLFHSRDRSDLEPSPLRQAKPIADLHAAQRFLSRYGWSGVWAA
WGPSPEGPPETPKGAALAEAVRRFQRANALPASGELDAATLAAMNRPRCGVPDMRPPPPSAPPSPPGPP
PRARSRRSPRAPLSLSRRGWQPRGYPDGGAAQAFSKRTLSWRLLGEALSSQLSAADQRRIVALAFRMW
SEVTPLDFREDLAAPGAAVDIKLGFGRRRHLGCPRAFDGSGQEFAHAWRLGDIHFDDDEHFTPPTSDTG
ISLLKVAVHEIGHVLGLPHTYRTGSIMQPNYIPQEPAFELDWSDRKAIQKLYGSCEGSFDTAFDWIRKER
NQYGEVMVRFSTYFFRNSWYWLYENRNNRTRYGDPIQILTGWPGIPTHNIDAFVHIWTWKRDERYFFQ
GNQYWRYDSDKDQALTEDEQGKSYPKLISEGFPGIPSPLDTAFYDRRQKLIYFFKESLVFAFDVNRNRV
LNSYPKRITEVFPAVIPQNHPFRNIDSAYYSYAYNSIFFFKGNAYWKVVNDKDKQQNSWLPANGLFPK
KFISEKWFDVCDVHISTLNM

>SGPr_069_SEQ ID NO:63
MVESAGRAGQKRPGFLEGGLLLLLLLVTAALVALGVLYADRRGIPEAQEVSEVCTTPGCVIAAARILQ
NMDPTTEPCDDFYQFACGGWLRRHVIPETNSRYSIFDVLRDELEVILKAVLENSTAKDRPAVEKARTLY
RSCMNQSVIEKRGSQPLLDILEVVGGWPVAMDRWNETVGLEWELERQLALMNSQFNRRVLIDLFIWN
DDQNSSRHIIYIDQPTLGMPSREYYFNGGSNRKVREAYLQFMVSVATLLREDANLPRDSCLVQEDMVQ
VLELETQLAKATVPQEERHDVIALYHRMGLEELQSQFGLKGFNWTLFIQTVLSSVKIKLLPDEEVVVYG
IPYLQNLENIIDTYSARTIQNYLVWRLVLDRIGSLSQRFKDTRVNYRKALFGTMVEEVRWRECVGYVN
SNMENAVGSLYVREAFPGDSKSMVELIDKVRTVFVETLDELGWMDEESKKKAQEKAMSIREQIGHPD
YILEEMNRRLDEEYSNVNFSEDLYFENSLQNLKVGAQRSLRKLREKVDPNLIIGAAVVNAFYSPNRNQI
VFPAGILQPPFFSKEQPQALNFGGIGMVIGHEITHGFDDNGGRNFDKNGNMMDWWSNFSTQHFREQSE
CMIYQYGNYSWDLADEQNVNGFNTLGENIADNGGVRQAYKAYLKWMAEGGKDQQLPGLDLTHEQL
FFINYAQVWCGSYRPEFAIQSIKTDVHSPLKYRVLGSLQNLAAFADTFHCARGTPMHPKERCRVW

>SGPr_212_SEQ ID NO:64
MRLKLKGSHLSAEVKAKYSQREGIAVNCCDVCDVHLKSLCECNYTGWHTLMSALDPHKPLAWALRP
FSPFLLTSSPALEAAGSPSQSPPWQIVNRLGHASSPVESGSEAGTTEASPTLGCVQERGTKGFRLEEGAG
AESSACKCVGESVDIHHFTPDEGKRRQAMNLRGVERHLLEPAVAAASSQGRQVLGRSTHSKMGRAGP
RRLLYLHKWALVRLPHWDRRAGRSPDSGGFFFMNSLRAISQSSTRGSFLAGVRPPVSSILTGGNHLCGT
RLCHEIAHAWFGLAIGARDWTEEWLSEGFATHLEDVFWATAQQLGLAFHTLAVDPAVCTSVSPATWS
PVRRGHMIDTEKALGSESDRLPVLALPFVGSVSIDSSTKFETFPEQVRQADLSLQVRDWAVAGPGECLP
QTVQGVGECPVGQGWPRAAFSLRSHMAFPLCMQRERRDAMLPRGDAGVKKLLQDLQQEGGMICSVF
GRCCSAAVWRAPQAADGKPGERLQPCSSPCKRPWSACDRCKTQTYLKCVLAVERAGLWLIECGEEEN
ECIQNDFEVFELDSWVDGDPICVMIFSSYSLDPQFSLRLLFLTVDAVSQPDEGAGLHGAYVQDHMAVE
RLGSKPSPSGHAPSPAGLTCASGAQMGTVGQSLHKGQISLPPLLQGLDLSSGGPIRNQIIMYQISLPPAHS
LNIHIASVVVVEKEGVGKGKGTSISVVAFGAKPSKDKTGHTSDSGASVIKHGLNPEKIFMQVHYLKGYF
LLRFLAKRLGDETYFSFLRKFVHTFHGQLILSQPSTEPLPSSHPANVCHIENVACFSVFSGEDFGPHLITF
QGSTPQPPLHATPREASEAAMPDVCDEYALSSRNWLSQPNSSFQSTESTHDAVPGSLDFIVHVAVGEEE
RSHVTGLPSTLQPRGALPFL

>SGPr_049_SEQ ID NO:65
MGPPSSSGFYVSRAVALLLAGLVAALLLALAVLAALYGHCERVPPSELPGLRDSEAESSPPLRQKPTPT
PKPSSARELAVTTTPSNWRPPGPWDQLRLPPWLVPLHYDLELWPQLRPDELPAGSLPFTGRVNITVRCT
VATSRLLLHSLFQDCERAEVRGPLSPGTGNATVGRVPVDDVWFALDTEYMVLELSEPLKPGSSYELQL
SFSGLVKEDLREGLFLNVYTDQGERRALLASQLEPTFARYVFPCFDEPALKATFNITMIHHPSYVALSN
MPKLGQSEKEDVNGSKWTVTTFSTTPHMPTYLVAFVICDYDHVNRTERGKEIRIWARKDAIANGSADF
ALNITGPIFSFLEDLFNISYSLPKTDIIALPSFDNHAMENWGLMIFDESGLLLEPKDQLTEKKTLISYVVSH
EIGHQWFGNLVTMNWWNNIWLNEGFASYFEFEVINYFNPKLPRNEIFFSNILHNILREDHALVTRAVA
MKVENFKTSEIQELFDIFTYSKGASMARMLSCFLNEHLFVSALKSYLKTFSYSNAEQDDLWRHFQMAI
DDQSTVILPATIKNIMDSWTHQSGFPVITLNVSTGVMKQEPFYLENIKNRTLLTSNDTWIVPILWIKNGT
TQPLVWLDQSSKVFPEMQVSDSDHDWVILNLNMTGYYRVNYDKLGWKKLNQQLEKDPKAIPVIHRL
QFIDDAFSLSKNNYIEIETALELTKYLAEEDEIIVWHTVLVNLVTRDLVSEVNIYDIYSLLKRYLLKRLNL
IWNIYSTIIRENVLALQDDYLALISLEKLFVTACWLGLEDCLQLSKELFAKWVDHPENEIPYPIKDVVLC
YGIALGSDKEWDILLNTYTNTTNKEEKIQLAYAMSCSKDPWILNRYMEYAISTSPFTSNETNIIEVVASS
EVGRYVAKDFLVNNWQAVSKRYGTQSLINLIYTIGRTVTTDLQIVELQQFFSNMLEEHQRIRVHANLQT
IKNENLKNKKLSARIAAWLRRNT

Figure 2I

>SGPr_026_SEQ ID NO:66
MASGEHSPGSGAARRPLHSAQAVDVASASNFRAFELLHLHLDLRAEFGPPGPGAGSRGLSGTAVLDLR
CLEPEGAAELRLDSHPCLEVTAAALRRERPGSEEPPAEPVSFYTQPFSHYGQALCVSFPQPCRAAERLQ
VLLTYRVGEGPGVCWLAPEQTAGKKKPFVYTQGQAVLNRAFFPCFDTPAVKYKYSALIEVPDGFTAV
MSASTWEKRGPNKFFFQMCQPIPSYLIALAIGDLVSAEVGPRSRVWAEPCLIDAAKEEYNGVIEEFLAT
GEKLFGPYVWGRYDLLFMPPSFPFGGMENPCLTFVTPCLLAGDRSLADVIIHEISHSWFGNLVTNANW
GEFWLNEGFTMYAQRRISTILFGAAYTCLEAATGRALLRQHMDITGEENPLNKLRVKIEPGVDPDDTY
NETPYEKGFCFVSYLAHLVGDQDQFDSFLKAYVHEFKFRSILADDFLDFYLEYFPELKKKRVDIIPGFEF
DRWLNTPGWPPYLPDLSPGDSLMKPAEELAQLWAAEELDMKAIEAVAISPWKTYQLVYFLDKILQKSP
LPPGNVKKLGDTYPSISNARNAELRLRWGQIVLKNDHQEDFWKVKEFLHNQGKQKYTLPLYHAMMG
GSEVAQTLAKETFASTASQLHSNVVNYVQQIVAPKGS

>SGPr_203_SEQ ID NO:67
MAAQCCCRQAPGAEAAPVRPPPEPPPALDVASASSAQLFRLRHLQLGLELRPEARELAGCLVLELCAL
RPAPRALVLDAHPALRLHSAAFRRAPAATRTPCAFAFSAPGPGPAPPPPLPAFPEAPGSEPACCPLAFRV
DPFTDYGSSLTVTLPPELQAHQPFQVILRYTSTDAPAIWWLDPELTYGCAKPFVFTQGHSVCNRSFFPCF
DTPAVKCTYSAVVKAPSGVQVLMSATRSAYMEEEGVFHFHMEHPVPAYLVALVAGDLKPADIGPRSR
VWAEPCLLPTATSKLSGAVEQWLSAAERLYGPYMWGRYDIVFLPPSFPIVAMENPCLTFIISSILESDEFL
VIDVIHEVAHSWFGNAVTNATWEEMWLSEGLATYAQRRITTETYGAAFTCLETAFRLDALHRQMKLL
GEDSPVSKLQVKLEPGVNPSHLMNLFTYEKGYCFVYYLSQLCGDPQRFDDFLRAYVEKYKFTSVVAQ
DLLDSFLSFFPELKEQSVDCRAGLEFERWLNATGPPLAEPDLSQGSSLTRPVEALFQLWTAEPLDQAAA
SASAIDISKWRTFQTALFLDRLLDGSPLPQEVVMSLSKCYSSLLDSMNAEIRIRWLQIVVRNDYYPDLHR
VRRFLESQMSRMYTIPLYEDLCTGALKSFALEVFYQTQGRLHPNLRRAIQQILSQGLGSSTEPASEPSTE
LGKAEADTDSDAQALLLGDEAPSSAISLRDVNVSA

>SGPr_157_SEQ ID NO:68
MDPKLGRMAASLLAVLLLLLERGMFSSPSPPPALLEKVFQYIDLHQDEFVQTLKEWVAIESDSVQPVPR
FRQELFRMMAVAADTLQRLGARVASVDMGPQQLPDGQSLPIPPVILAELGSDPTKGTVCFYGHLDVQP
ADRGDGWLTDPYVLTEVDGKLYGRGATDNKGPVLAWINAVSAFRALEQDLPVNIKFIIEGMEEAGSV
ALEELVEKEKDRFFSGVDYIVISDNLWISQRKPAITYGTRGNSYFMVEVKCRDQDFHSGTFGGILHEPM
ADLVALLGSLVDSSGHILVPGIYDEVVPLTEEEINTYKAIHLDLEEYRNSSRVEKFLFDTKEEILMHLWR
YPSLSIHGIEGAFDEPGTKTVIPGRVIGKFSIRLVPHMNVSAVEKQVTRHLEDVFSKRNSSNKMVVSMTL
GLHPWIANIDDTQYLAAKRAIRTVFGTEPDMIRDGSTIPIAKMFQEIVHKSVVLIPLGAVDDGEHSQNEK
INRWNYIEGTKLFAAFFLEMAQLH

>SGPr_154_SEQ ID NO:69
MAQRCVCVLALVAMLLLVFPTVSRSMGPRSGEYQRASRIPSQFSKEERVAMKEALKGAIQIPTVTFSSE
KSNTTALAEFGKYIRKVFPTVVSTSFIQHEVVEEYSHLFTIQGSDPSLQPYLLMAHFDVVPAPEEGWEVP
PFSGLERDGVIYGRGTLDDKNSVMALLQALELLLIRKYIPRRSFFISLGHDEESSGTGAQRISALLQSRGV
QLAFIVDEGGFILDDFIPNFKKPIALIAVSEKGSMNLMLQVNMTSGHSSAPPKETSIGILAAAVSRLEQTP
MPIIFGSGTVVTVLQQLANEVYGEKSLNQCNNQDHHGTHHIQSRVAQATVNFRIHPGQTVQEVLELTK
NIVADNRVQFHVLSAFDPLPVSPSDDKALGYQLLRQTVQSVFPEVNITAPVTSIGNTDSRFFTNLTTGIY
RFYPIYIQPEDFKRIHGVNEKISVQAYETQVKFIFELIQNADTDQEPVSHLHKL

>SGPr_088_SEQ ID NO:70
MAALTTLFKYIDENQDRYIKKLAKWVAIQSVSAWPEKRGEIRRMMEVAAADVKQLGGSVELVDIGKQ
KLPDGSEIPLPPILLGRLGSDPQKKTVCIYGHLDVQPAALEDGWDSEPFTLVERDGKLYGRGSTDDKGP
VAGWINALEAYQKTGQEIPVNVRFCLEGMEESGSEGLDELIFARKDTFFKDVDYVCISDNYWLGKKKP
CITYGLRGICYFFIEVECSNKDLHSGVYGGSVHEAMTDLILLMGSLVDKRGNILIPGINEAVAAVTEEEH
KLYDDIDFDIEEFAKDVGAQILLHSHKKDILMHRWRYPSLSLHGIEGAFSGSGAKTVIPRKVVGKFSIRL
VPNMTPEVVGEQVTSYLTKKFAELRSPNEFKVYMGHGGKPWVSDFSHPHYLAGRRAMKTVFGVEPD
LTREGGSIPVTLTFQEATGKNVMLLPVGSADDGAHSQNEKLNRYNYIEGTKMLAAYLYEVSQLKD